11-2022

More Than Yes and No: Predicting the Magnitude of Non-Invariance Between Countries from Systematic Features

Johannes A. Karl
Victoria University of Wellington

Ronald Fischer
Victoria University of Wellington

Abstract

Measurement Invariance has long been the cornerstone of cross-cultural comparisons. Nevertheless, over time a research tradition has developed in which invariance tests are applied with the stated end goal of finding invariance between measures and an implicit view that non-invariance is a barrier to cross-cultural research. In the current paper we aim to challenge this view and urge researchers to consider non-invariance critically not as barrier, but as opportunity for cross-cultural research. Specifically, we show how invariance effect sizes of items can be used to understand psychometric distances between countries and formulate novel hypotheses on cultural differences. Using a previously published dataset on the cross-cultural comparability of subjective happiness from 59 countries, we show how invariance effect sizes can be used to detect problematic items and variables which shape the psychometric similarity of countries. Focusing on item differences, we showed that negatively worded items are performing markedly worse in cross-cultural comparisons and that this effect is exacerbated if countries are linguistically distant. Additionally, we showed that country level variables such as GDP or environmental factors such as temperature can be used to cluster similarities in psychometric functioning, creating novel possibilities to systematize sources of non-invariance on a granular level.

Keywords: invariance; equivalence; dMacs; MGCFA; linguistic distance

More Than Yes and No: Predicting the Magnitude of Non-Invariance Between Countries from Systematic Features

Testing the cross-cultural comparability of instruments is an essential part of cross-cultural research. Over the last four decades, the frameworks and statistical techniques to determine whether and how data patterns can be compared across populations have significantly matured and many excellent summaries of the conceptual and procedural steps are now available (Fischer & Karl, 2019; Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000). In recent years, awareness of issues surrounding group comparisons has also started to rise in fields outside cross-cultural research and researchers focusing on identity, gender, and ideology have started to adopt these methods (Brandt et al., 2021). Nevertheless, even within journals that explicitly focus on cross-cultural research, the minimum steps to ensure that data can be compared are often not reported (Boer et al., 2018). One reason for this might be that researchers might perceive tests of invariance as gatekeepers to meaningful research because a decision that item bias has been found often precludes answering the questions of interest for researchers. Given the difficulty of attaining levels of invariance in real data that would allow a straightforward comparison of data across populations, researchers may be reluctant to conduct and report those tests (Boer et al., 2018).

This view of invariance may be partially driven by an all-or-nothing mentality within traditional invariance testing frameworks. Invariance is typically treated as a dichotomous category, with data either showing levels of invariance that are above or below a commonly accepted threshold indicating sufficient data similarity (Welzel et al., 2021). Yet, it may be more productive to think about invariance as a continuous property of data, which then becomes amenable to further inquiry and may actually contribute novel insights in both cultural and substantive psychological processes (Fischer et al., 2022; Meuleman et al., 2022). This shift in the conceptualization of invariance parallels recent attempts to move beyond binary significance statements and rather focus on the magnitude of cultural differences (Matsumoto et al., 2001). Therefore, our first goal is to expand the perspective on invariance testing by explicitly focusing on the effect size of invariance parameters. We provide an example using happiness data that demonstrates how effect sizes of invariance parameters can be used in different ways to provide further insights into cross-cultural data.

Our second goal is to focus on linguistic similarity as a largely ignored problem in cross-cultural research. Researchers pay attention to translation methods, yet the linguistic similarity of the languages being used may systematically shape response patterns. By drawing upon available linguistic data sets, we demonstrate that the extent of invariance in a specific happiness scale is partially explicable by linguistic (dis)similarity. Our point is that by focusing on effect sizes in invariance estimates, we can start to explore additional factors, including linguistic similarity, as potential contributors of both bias and substantive variance in psychological responses.

We will start by briefly reviewing classic frameworks of invariance and their rationale for making decisions on bias and equivalence. We then present one promising effect size

parameter for invariance testing and briefly outline how effect sizes can provide novel insights for cross-cultural psychologists. Finally, we report about research in the domain of happiness and discuss the promise of linguistic distance metrics for cross-cultural research before we finish the paper with some conclusions.

Cross-Cultural Equivalence and Bias Frameworks

The issue of examining data quality has been a central topic for psychologists since the beginning of psychometric testing. In the last four decades, researchers have made significant advances in describing possible biases in cross-cultural data and specified a level of hierarchies of equivalence, typically within a latent variable framework (e.g. Fischer & Karl, 2019; Fischer & Smith, 2021; Fontaine, 2005; Lubke et al., 2003; Messick, 1991; Van de Vijver & Leung, 2021; Vandenberg & Lance, 2000). Here, we use the framework championed by Fontaine (2005), which differentiates four levels of equivalence[1] in order to address three fundamental questions: (1) do the same theoretical constructs account for observed test behavior, (2) can we use the same observed variables to measure our theoretical constructs of interest across different groups, and (3) what type of inferences can we draw from the observed scores across cultural groups?

Functional Equivalence

The most basic and fundamental level of equivalence is functional equivalence. The first question to address this level is to ask whether a specific construct (e.g., happiness) can be expected to be psychologically relevant in another culture. This issue needs to be addressed prior to any measurement development or data collection. The most appropriate methods to tackle this level are extensive theoretical and conceptual analyses and via qualitative (and possibly culture-specific quantitative) studies in each cultural group separately. The main epistemological question is whether the same construct can be assumed to account for behavioral differences in each group. For example, we may ask whether the concept of happiness exists in different cultural groups and how this concept may function psychologically – what are the mental representations of happiness, how does happiness influence daily functioning, what correlates of happiness could we expect to find in each of the cultural contexts?

Structural Equivalence

The next higher level after having established (or better: proposed functional equivalence) is structural equivalence. It is concerned with the question whether the same observed variables or items can be used to measure the same underlying theoretical variable in each of the cultural groups. A number of researchers have combined functional and structural

[1] While philosophical differences exist between the two terms, due to pragmatic similarities and in line with previous literature we use the terms equivalence and invariance interchangeably.

equivalence under construct equivalence (Van de Vijver & Leung, 2021). Here, we follow Fontaine (for a historical overview of MGCFA see: Byrne & Matsumoto, 2021; Fontaine, 2005) in separating them because for theoretical and operational reasons it makes sense to separate the theoretical and ethnographic focus of functional equivalence from the operational concerns of construct equivalence. It is important to emphasize that functional equivalence is a prerequisite for structural equivalence. If a researcher decides to declare a construct absent or qualitatively different in one cultural context compared to another, then no group comparisons are possible. It is of course possible to continue emic research to provide a rich in-depth understanding within each of the contexts. If functional equivalence is assumed, it becomes possible to consider measuring the concept and identify relevant and representative indicators within each context. The important question to be addressed with tests of structural equivalence is whether the items or indicators are relevant and representative in each context, as mentioned before. Typically, individuals are presented with a small set of stimuli (often questionnaire items) drawn from a potential pool of stimuli that could represent the theoretical construct and the response to these stimuli that are assumed to provide some information on the particular theoretical variable of interest. Psychologists are often interested in generalizing from these observed stimuli responses within a specific testing situation to broader and presumably stable characteristic of the participant. Therefore, it becomes important to examine whether this small subset of items is relevant and representative for providing information about the theoretical construct across cultures. Naturally, irrelevant items would measure some other theoretical construct, which introduces systematic error in the measurement. Non-representative items would capture behavior that is not reflecting the core aspects of the domain of interest. Again, systematic error is introduced by the inclusion of such items. To give some fictitious examples, if a specific culture considers negative feelings to be part of a cycle of happiness, then excluding those items would lead to underrepresentation of the construct.

Tests of structural equivalence rely on the proposition that items should have a non-trivial weight parameter within each of the cultural groups. This implies similar internal structures, which can be tested via internal consistency tests such as Cronbach's alpha or structure-oriented tests including Confirmatory Factory Analysis [CFA], Exploratory Factor Analysis [EFA], or Multidimensional Scaling [MDS]. The assumption of structural equivalence is considered met if the association of the item with the presumed construct is above a threshold, either indicated by a significant item-total correlation or by a factor loading above a certain threshold (such as 0.30 J. W. Osborne et al., 2008).

Metric Equivalence

However, this assessment does not indicate how similar these loadings or parameter weights are. This is the focus of the next higher level of equivalence, which is commonly called metric equivalence (Fontaine, 2005) or measurement unit equivalence (van de Vijver & Leung, 2021). The important issue at this point is whether the measurement units are identical across groups and empirically comparable weight parameters (e.g., factor loadings) are estimated in each cultural group. Statistically, this is typically done by demonstrating that

loadings of the items on the underlying factors or the location of items in a specific multidimensional space are not sufficiently distinct at a statistical level across cultural groups. If a test suggests that the loadings are *not* statistically different from each other, the researcher can compare patterns of scores across cultural or ethnic groups. As is implicated in the previous sentence, the question typically becomes a binary decision, with an item or more often combinations of items either being above or below a certain threshold.

If items are above the desired threshold and metric equivalence of a scale can be assumed within the specific samples, then it is possible to draw conclusions about correlations and score patterns. For example, we may compare correlations between happiness and demographic variables across cultural samples. However, it is not yet possible to directly compare the scores and interpret them in terms of cultural differences in happiness, because other biases within the data have not been ruled out yet. Again, typical tests for assessing metric equivalence are various types of factor analysis as well as logistic regression (Fischer & Karl, 2019)

Scalar Equivalence

Direct comparison of scores is only possible if full-score equivalence (Fontaine, 2005) or scalar equivalence (van de Vijver & Leung, 2021) is being met. In this case, individuals with the same score on a specific test are assumed to share the same underlying latent score, independent of cultural context and, importantly, differences in observed scores reflect 'true' differences in the proposed theoretical variable. Statistically, this question is being answered by examining the equality of the intercepts of the factor loadings or random parameters or thresholds in various types of item response theory models. Only in the absence of intercept or threshold differences can any observed score differences be validly interpreted as reflecting substantive differences in the proposed underlying theoretical construct. Again, these decisions are based on fit indicators falling above or below a certain threshold, the question of equivalence again becoming a binary decision.

As can be imagined, researchers may feel uncomfortable to rely on these binary decision-criteria and in the likely case that a test does not meet these standards, the research project is typically considered finished. However, we argue that psychometric non-invariance should be viewed not as an obstacle to meaningful cross-cultural research but should be seen as a rich source of data to investigate cultural similarities and differences, allowing a much richer insight into the concept of 'culture'. Admittingly, this is not a novel thought and others have succinctly expressed similar concerns in the past for example: "From our perspective, measurement non-invariance is not a showstopper, but rather an outcome to be explained. Non-invariance provides analysts with an opportunity to more closely consider sources of variation and how such variation maps onto measurement—and through such explorations come conceptual and theoretical development." (Medina et al., 2009, p. 358).

How to Run Invariance Analyses

As indicated above, functional equivalence is typically addressed using qualitative and conceptual tools. Structural, metric, and scalar invariance can be tested with several different programs. The most common strategy is to use either a structural equation modelling program or a program dedicated to a variation of item response theory. Commercial programs such as MPlus, AMOS, EQS, WINLOG. Xcaliber, or RASCAL can be costly. Fortunately, high quality open-source alternatives that do not require advanced programming skills are available today. The most promising alternatives are JASP (using structural equation modelling, see jasp-stats.org) and various packages available via R that allow both structural equation and item response options. For a tutorial that describes the step-by-step process for testing invariance with both structural equation modelling and item response theory in R, please see Fischer and Karl (2019).

Non-Invariance as Continuum

Specifically, the shift from a binary yes-no criterion towards a focus on the effect sizes in equivalence testing can open exciting new opportunities for exploring cross-cultural differences in psychological processes more broadly and sources of cross-national invariance in items and constructs specifically. Recent advancements in quantifying the degrees of invariance of items between groups open interesting new research avenues about the potential sources of invariance on an item level. One such advancement is the introduction of effect sizes that quantify the degree of invariance of items between groups (Gunn et al., 2019; Nye & Drasgow, 2011). These effect sizes do not only allow researchers to get a finer grained perspective on metric invariance in their data but can themselves become targets of insightful cross-cultural research.

Effect Sizes for Non-Invariance (dMacs)

Nye and Drasgow (2011) first proposed an effect size equivalent measure for differences in mean and covariance structures (called dMACS). This index is calculating the degree of non-equivalence between two groups per individual item in relation to the item variability and can be interpreted in a similar way as established effect sizes like Cohen's *d* or *r* (Cohen, 1988). These estimates can be calculated for both factor loadings and factor intercepts. Values of smaller than 0.20 are being considered small, values of about 0.50 are medium, and 0.80 or greater are considered large. These values could be used as input into binary decisions, such as when deciding a specific criterion that a researcher is accepting as trivial and any items that show differences above this threshold need to be excluded (or set to varying, as in the case of partial invariance, see Byrne et al., 1989; Shi et al., 2019). However, the true power we believe lies in the empirical estimation of the size of the invariance of the parameters of interest.

Currently, the index is only available for unidimensional scales and can only be applied for pairwise comparisons of scores between two populations. The need to compute all

pairwise comparisons is nevertheless not so much an issue as the index is based on a) effect sizes and not significance, which should not be affect by the number of comparisons, and b) this pairwise comparison opens interesting opportunities for clustering of distant samples.

One barrier to a wider spread adoption of these effect sizes is the limited implementation in commonly used statistical programs (Gunn et al., 2019). In our current article, we aim to address this by providing an applied example of an implementation of these effect sizes in the R language (R Core Team, 2018). We highlight the possibility of using continuous indicators of non-invariance as a basis for substantiative research using previously published datasets. Specifically, we focus on happiness, which has been of significant interest for cross-cultural research. A number of studies (e.g. Tsai & Park, 2014) have suggested that cultures may differ in how they value and interpret happiness, making a more focused analysis of the structure of widely used happiness scales informative. By focusing on the empirical extent of item biases, we may be able to provide new insights into the psychological functioning of happiness across cultures.

One area of particular interest is the linguistic similarity of languages that individuals are using for answering questions on happiness. Studies that used the geographical proximity of languages have found that closer proximity is systematically linked to the co-lexification of emotion terms, indicating that language features might provide a scaffold for cultural similarity and differences (Jackson et al., 2019). Language comparisons have often been limited to a small range of well-studied languages. Moving to a broader range of comparisons, Jaeger (2019) recently produced Pointwise Mutual Information scores between sound classes of languages from phonetic transcriptions of word lists of more than 7000 languages present in the Automated Similarity Judgment Program. This approach, therefore, opens new opportunities for a more systematic analysis of linguistic similarity effects on invariance tests.

In summary, in this paper, we aim to make several major contributions. First, we apply the idea of effect sizes for invariance parameters to demonstrate that this provides useful novel information beyond the previous dichotomous treatment of invariance parameters. We also demonstrate how the degree of invariance can be utilized for further analyses of item wording effects as well as network analyses that can provide further insights. Second, we use a happiness variable as a test case to show these effects. By focusing on the invariance of a happiness index, we are contributing to previous discussions on the conceptualization of subjective wellbeing and in particular happiness from an empirical perspective (for discussions focused on the cultural construction of the construct see: Uchida et al., 2004). Third, by using linguistic distance as a predictor variable of the invariance parameters, we are testing whether a largely unexplored major confound may contribute to explaining variability in psychological response patterns. We also test the relevance of national wealth and temperature, as these variables have been shown to influence wellbeing globally (Fischer & Van de Vliert, 2011).

Methods

Participants

We used previously published data from the International Situations Project, available on the Open Science Framework (https://osf.io/jrbt3/), selecting only countries with N > 100 to allow for a robust convergence of the CFA model. This left us with data from 59 countries with 15,097 participants (see Table 1 for descriptive information). This data has previously been published (Gardiner et al., 2020) and is used here to illustrate the benefits of continuous assessments of measurement invariance.

Table 1

Sample Descriptive Information

Country	N	Age (Mean)	Age (SD)	Female %
Argentina	140	24.279	5.658	78.571
Australia	196	19.837	3.581	76.020
Austria	113	21.257	2.367	81.416
Bolivia	135	21.015	2.158	57.778
Brazil	310	23.690	7.091	71.935
Bulgaria	152	25.020	6.458	69.737
Canada	304	21.849	3.966	78.618
Chile	386	21.469	3.083	66.062
China	432	22.627	4.372	47.917
Colombia	181	21.680	4.160	74.033
Croatia	218	21.459	1.696	64.679
Czech Republic	193	22.648	4.820	80.829
Denmark	246	22.923	5.102	79.268
Estonia	293	25.877	7.669	83.959
France	231	22.580	6.275	84.416
Georgia	140	20.293	1.789	80.000
Germany	458	24.356	6.367	74.454
Greece	225	22.569	5.284	80.000
Hong Kong	144	18.993	1.260	58.333
Hungary	178	21.764	2.072	59.551
India	221	22.376	4.650	49.774
Indonesia	131	21.832	5.066	51.908
Israel	173	25.416	4.286	60.694
Italy	717	21.862	3.730	64.575
Japan	243	22.564	4.822	61.728
Jordan	141	19.865	2.135	80.851
Kenya	139	21.165	1.898	65.468
Latvia	169	24.870	6.090	82.840

Table 1 continued

Lithuania	145	20.262	1.748	77.931
Malaysia	230	21.517	2.794	70.435
Mexico	247	23.850	6.068	57.895
Netherlands	301	20.136	3.028	81.063
New Zealand	129	19.194	4.430	86.047
Nigeria	135	24.719	5.660	33.333
Norway	159	23.887	5.039	74.214
Pakistan	114	20.614	2.735	50.000
Palestine	295	22.173	4.809	83.390
Philippines	337	19.694	2.206	67.953
Poland	234	22.346	5.322	83.333
Portugal	157	21.771	5.980	87.261
Romania	177	22.836	5.572	57.062
Russia	159	21.899	4.701	77.987
Senegal	635	23.315	2.249	47.402
Serbia	185	19.724	1.257	85.946
Singapore	136	20.926	2.128	77.941
Slovakia	148	22.405	2.713	69.595
Slovenia	123	20.585	2.336	56.911
South Africa	256	22.199	4.741	66.406
South Korea	281	22.345	2.248	58.363
Spain	419	19.730	3.467	85.203
Sweden	130	28.333	1.155	70.000
Switzerland	755	22.351	4.851	83.709
Taiwan	162	19.710	1.345	76.543
Thailand	196	19.265	1.155	77.041
Turkey	329	21.085	2.799	68.085
Ukraine	244	20.619	1.911	77.049
United Kingdom	136	25.640	8.080	88.971
United States	1366	19.857	3.118	67.423
Vietnam	168	19.048	1.326	76.786

Measures

The subjective happiness scale (Lyubomirsky & Lepper, 1999) is a four-item measure of global happiness measured on a Likert-scale from 1 to 6. The items were: “In general, I consider myself:” (1 not a very happy person – 7 a very happy person); “Compared to most of my peers, I consider myself:” (1 less happy – 7 more happy); “Some people are generally very happy. They enjoy life regardless of what is going on, getting the most out of everything. To what extent does this characterization describe you?” (1 not at all– 7 a great deal); “Some people are generally not very happy. Although they are not depressed, they never seem as

happy as they might be. To what extent does this characterization describe you?" (1 not at all– 7 a great deal).This measure has shown at least metric equivalence across a range of countries in previous studies (Zager Kocjan et al., 2021). See Table 2 for reliability information.

Table 2

Reliability Information per Country

Country	α	ω	GLB	H
Argentina	.803[.755, .850]	.808[.755, .861]	.826	.837
Australia	.850[.818, .883]	.852[.818, .887]	.882	.905
Austria	.840[.792, .888]	.848[.802, .894]	.867	.867
Bolivia	.852[.814, .890]	.852[.811, .893]	.889	.892
Brazil	.842[.814, .871]	.846[.818, .874]	.851	.873
Bulgaria	.886[.856, .916]	.885[.855, .916]	.922	.907
Canada	.873[.851, .895]	.874[.851, .897]	.871	.888
Chile	.870[.850, .891]	.871[.849, .892]	.884	.893
China	.781[.749, .813]	.795[.764, .827]	.844	.88
Colombia	.677[.610, .744]	.696[.626, .766]	.802	.856
Croatia	.879[.854, .904]	.879[.853, .906]	.909	.898
Czech Republic	.867[.838, .897]	.866[.834, .897]	.893	.903
Denmark	.889[.867, .911]	.889[.866, .912]	.898	.907
Estonia	.862[.838, .886]	.862[.836, .888]	.896	.899
France	.837[.805, .869]	.836[.800, .871]	.889	.878
Georgia	.751[.690, .812]	.749[.682, .816]	.821	.857
Germany	.886[.869, .902]	.884[.866, .901]	.908	.912
Greece	.812[.776, .849]	.813[.772, .853]	.853	.867
Hong Kong	.782[.723, .841]	.803[.751, .855]	.844	.888
Hungary	.827[.788, .867]	.828[.787, .870]	.844	.854
India	.606[.524, .688]	.607[.523, .691]	.649	.648
Indonesia	.538[.418, .657]	.625[.533, .718]	.729	.995
Israel	.685[.612, .757]	.716[.649, .783]	.790	.819
Italy	.820[.800, .839]	.819[.797, .841]	.847	.868
Japan	.793[.754, .833]	.788[.744, .831]	.835	.874
Jordan	.656[.568, .744]	.696[.617, .776]	.756	.812
Kenya	.607[.517, .697]	.637[.543, .731]	.739	.789
Latvia	.882[.853, .911]	.879[.848, .909]	.918	.927
Lithuania	.859[.823, .896]	.867[.831, .903]	.89	.920
Malaysia	.590[.515, .665]	.637[.566, .709]	.769	.836
Mexico	.718[.666, .770]	.727[.673, .781]	.826	.868
Netherlands	.889[.871, .908]	.892[.871, .912]	.920	.903
New Zealand	.828[.782, .874]	.836[.789, .882]	.894	.908
Nigeria	.637[.545, .729]	.656[.570, .742]	.793	.986

Table 2 continued

Norway	.862[.829, .896]	.862[.826, .897]	.887	.891
Pakistan	.488[.347, .628]	.608[.512, .705]	.745	.995
Palestine	.620[.552, .687]	.666[.607, .725]	.742	.811
Philippines	.791[.757, .825]	.797[.762, .833]	.843	.862
Poland	.873[.847, .899]	.872[.845, .899]	.895	.907
Portugal	.841[.801, .882]	.840[.799, .881]	.880	.873
Romania	.802[.757, .846]	.814[.768, .859]	.878	.889
Russia	.854[.818, .890]	.854[.817, .892]	.871	.879
Senegal	.508[.449, .568]	.525[.468, .582]	.597	.647
Serbia	.850[.816, .885]	.848[.812, .884]	.905	.946
Singapore	.860[.821, .898]	.865[.828, .902]	.893	.927
Slovakia	.829[.786, .873]	.835[.791, .878]	.873	.879
Slovenia	.840[.796, .884]	.839[.793, .886]	.868	.884
South Africa	.854[.827, .882]	.857[.828, .886]	.881	.901
South Korea	.884[.862, .905]	.882[.859, .905]	.904	.920
Spain	.861[.841, .882]	.864[.842, .886]	.889	.873
Sweden	.893[.865, .922]	.893[.862, .924]	.902	.917
Switzerland	.844[.827, .862]	.843[.825, .862]	.870	.871
Taiwan	.866[.833, .900]	.874[.842, .906]	.916	.938
Thailand	.864[.834, .895]	.867[.836, .897]	.898	.902
Turkey	.851[.825, .877]	.851[.825, .878]	.872	.880
Ukraine	.760[.716, .805]	.766[.719, .814]	.844	.878
United Kingdom	.893[.864, .921]	.894[.865, .924]	.935	.914
United States	.835[.821, .848]	.838[.823, .852]	.867	.892
Vietnam	.654[.578, .730]	.677[.601, .753]	.796	.861

Linguistic Distance. To assess the linguistic difference between two countries, we used the aggregated pointwise mutual information (PMI) estimated by Jaeger (2019). The average linguistic distance in our dataset was .7882 (SD = .1434) with the minimum linguistic distance between Indonesian and Malay (.0439) and the maximum distance between Ukrainian and Japanese (.9160).

GDP. We used the Gross Domestic Product (GDP) in Purchase Power Parity per capita in international US$ averaged for the year 2019 (World Bank, 2020). GDP in the past has been found to strongly relate to well-being and life satisfaction (Deaton, 2008; for a nuanced discussion see: Stanca, 2010).

Temperature. We used the average yearly temperature at the capital city of each nation state (reported by Gardiner et al., 2020; taken from *Travel Weather Averages (Weatherbase)*, 2022). As previous work has shown, temperature is a psychologically important consequence of latitude (Van de Vliert, 2007; Van de Vliert & Van Lange, 2020) and might shape a wide range of psychological constructs (Fischer & Van de Vliert, 2011; Georgas et al., 2004).

Data Analysis

We computed the psychometric similarity between countries by fitting a confirmatory factor analysis model for the subjective happiness scale with lavaan (Rosseel, 2012) for each pairwise comparison of countries. We then extracted *dmacs* effect sizes for each combination of countries for both the factor loadings and intercepts. We refer to this score as psychometric distance. To ease interpretation of some analysis this score is presented as its inverse, and we will refer to this as psychometric similarity. For readers interested in reproducing the analysis, all the code and the underlying data are available on on the Open Science Framework (https://osf.io/7wrk2/)**.** A step-by-step tutorial for running invariance analyses, we refer the reader to detailed primer which explains how to compute these effect sizes (Fischer & Karl, 2019).

Results

The overall CFA model suggested acceptable fit in the total dataset: X^2 (2, *N* = 15,097) = 122.31, p < .01, CFI = .995, TLI = .985, RMSEA = 0.063, SRMR = 0.015. A test of the configural model showed still acceptable fit overall: CFI = .993, RMSEA = 0.075, SRMR = 0.015. A test for metric invariance showed a considerable drop in fit: CFI = .976, ΔX^2 .018, RMSEA = .09, ΔRMSEA = -.017. These values suggested substantive variation in loading parameters across samples. We therefore extracted dmacs scores for the factor loadings across all pairwise comparisons and further analyzed the effect sizes.

We initially investigated the role that item direction may play in the psychometric comparability across countries. Overall, we found a low average difference between the countries on the subjective happiness scale, but also that the single reverse coded item of the subjective happiness scale showed the highest average psychometric distance between countries. Importantly, it also showed the highest standard deviation, indicating the potential presence of clusters of countries which may be more or less dissimilar in responses to the negatively worded item. We show the results in Figure 1. Similarly, when taken together the three positive items showed a substantially lower psychometric distance compared to the negatively worded item (See Figures 1-2).

Next, we investigated whether the psychometric similarity between countries can be meaningfully predicted by other cultural distance indicators such as linguistic similarity. Given the previously noted differences in negative vs positively phrased items, we regressed the psychometric similarity for positive and negative items separately onto their pairwise linguistic difference score. Due to several countries speaking the same language we ran the analysis once with these country pairs included and a second time with these countries excluded. For both positively (B = -.0193, *p* = .0965) and negatively worded items (B = -.0931, *p* < .001), greater linguistic distance was related to lower psychometric similarity. This effect became more pronounced when excluding pairs of countries from the analysis that had a linguistic distance of zero: the relationship for both positive (B = -.0492, *p* < .01)

Figure 1.
Distribution of Pairwise Psychometric Distances by Items

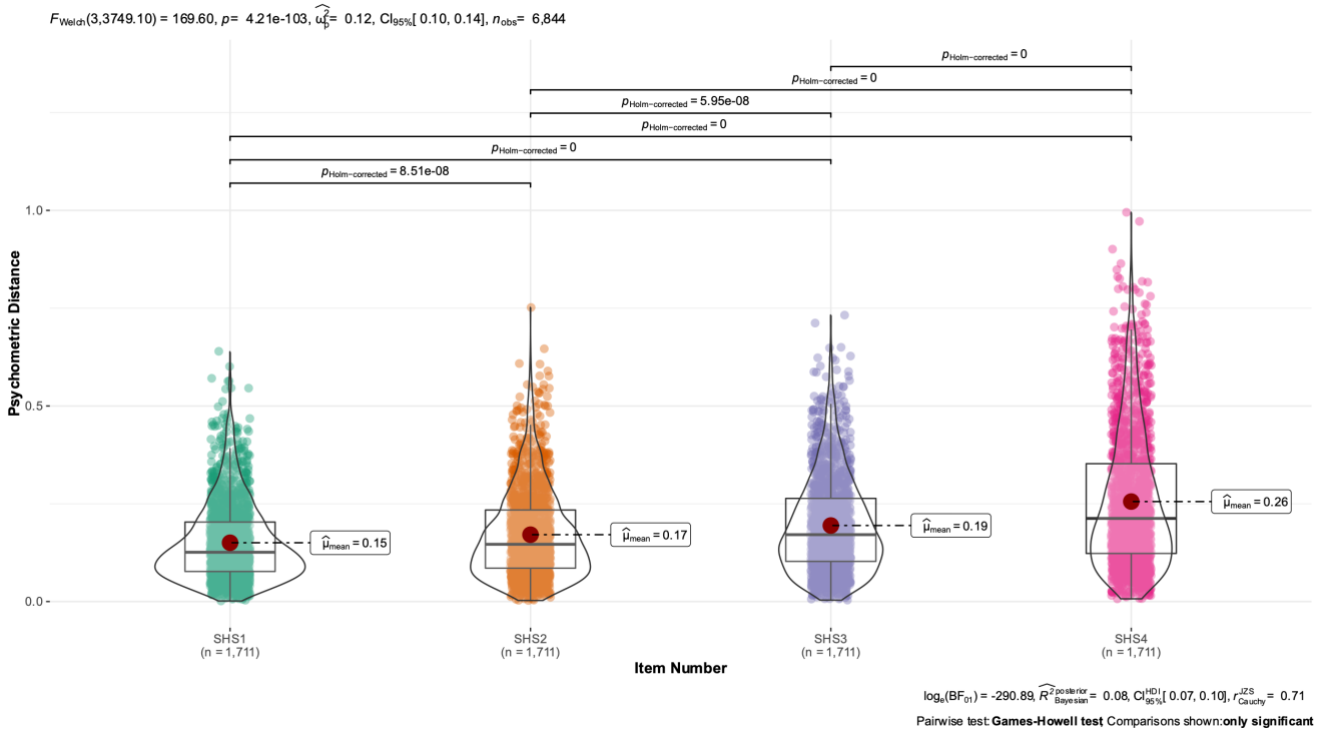
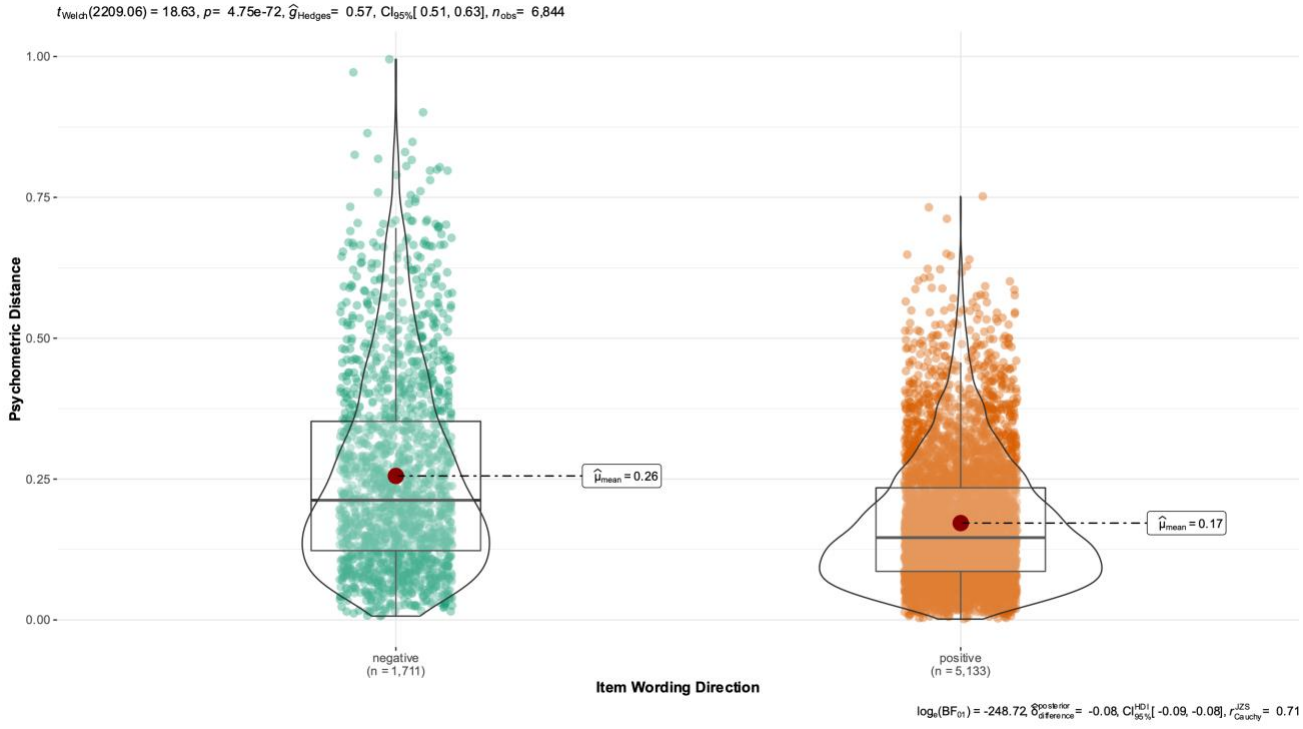


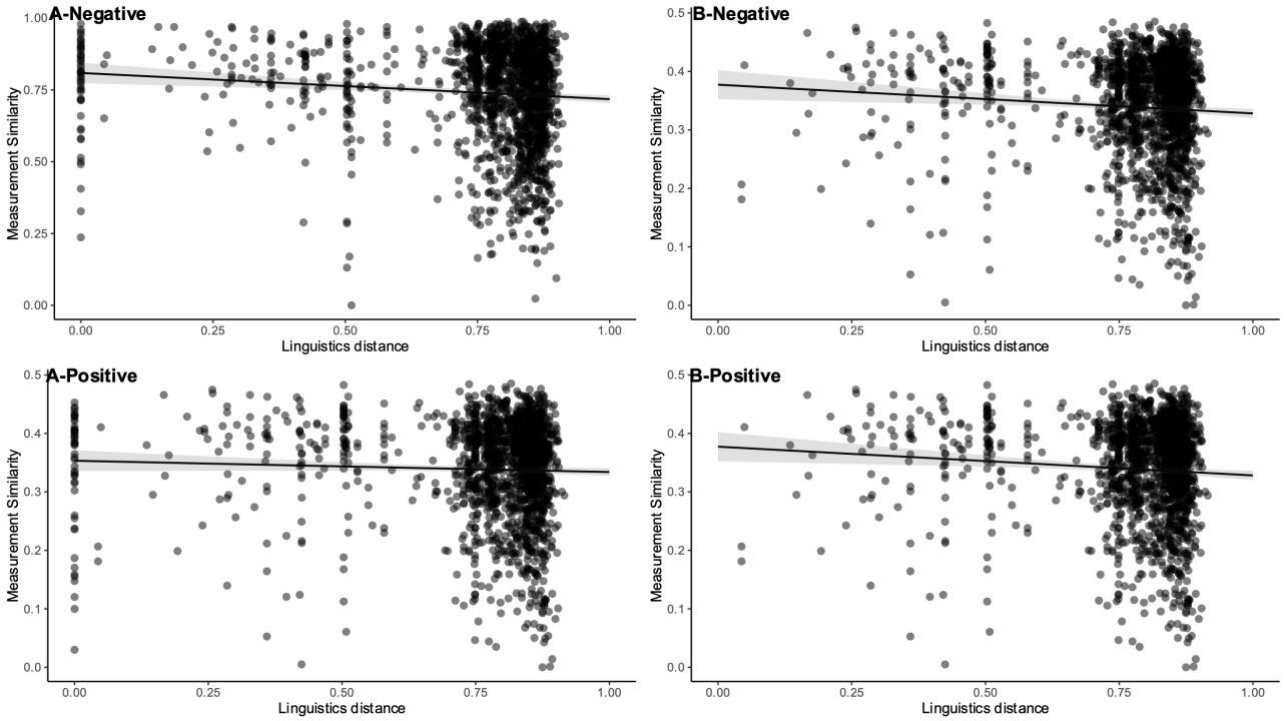
Figure 2.
Distribution of Pairwise Psychometric Distances Grouped by Item Direction



and negatively worded items (B = -.153, $p < .001$) was now highly significant. The effect was stronger for negatively compared to positively phrased items. We show the results in Figure 3.

Figure 3

Predicting Psychometric Similarity by Linguistic Distance Separated for Positive and Negative Items.



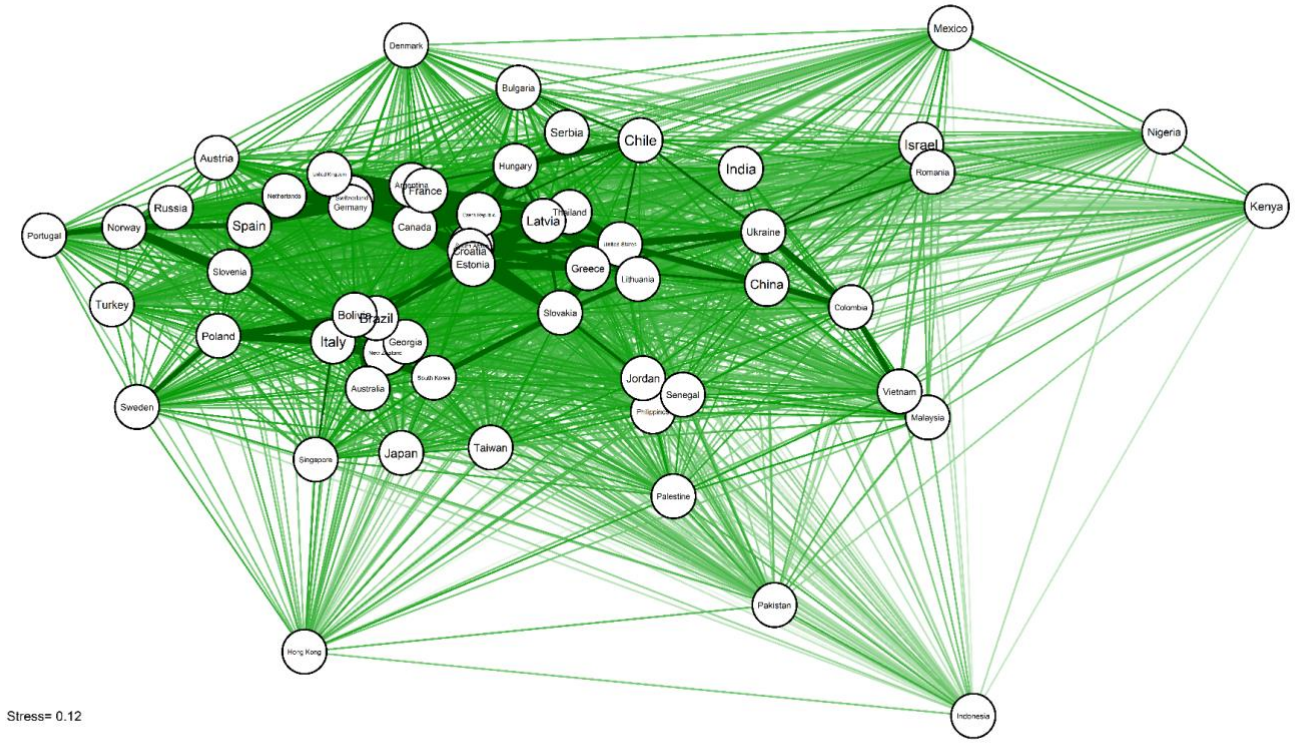
Note. Graph A- Including countries with zero linguistic distance, Graph B excluding countries with zero linguistic distance. Each dot represents a pairwise comparison, overlap therefore reflects density.

Finally, we investigated whether countries can be meaningfully clustered according to their average psychometric similarity across items. To examine this, we used a graphical network model to model the connection between countries as edge weights. We used a MDS procedure to extract the placement of countries along two axes (see Figure 4 for the two-dimensional solution). The MDS attempts to replicate the observed relationships of the network as accurately as possible while minimizing stress in a two-dimensional Cartesian system. The resulting axes can be interpreted by the researchers either conceptually via visual inspection or empirically by predicting countries position along the axes by other country level variables (for a general introduction see: Kruskal & Wish, 1978). To interpret the clustering solution, we first examined the role of GDP, considering the importance of GDP for clustering nations globally (Hofstede, 1980; Inglehart, 1997). Indeed, the first axis was strongly related to GDP with high GDP countries being placed substantially further to

the negative pole of dimension 1 ($r(53) = -.57$, $p < .001$). In contrast, GDP was unrelated to position of countries along the second axis ($r(53) = -.13$, $p = .34$). We next explored the potential to predict clustering along geographical features and found that average yearly temperature predicted both the first axis ($r(59) = .45$, $p < .001$) and the second axis ($r(59) = -.38$, $p < .001$). Finally, we also compared the congruence of the two-dimensional MDS solution for linguistic similarity with the two-dimensional MDS solution of the psychometric distance. Overall, we found low congruence ($\phi_1 = .25$, $\phi_2 = .32$). This implies that while linguistic effects are systematically influencing pairwise country comparisons, the overall network of psychometric distances is not reducible to linguistic distances and is likely shaped by a wide range of culture level similarities and differences such as economic or environmental factors (as demonstrated in our analyses).

Figure 4

MDS Adjusted Network Graph Based on Pairwise Country Similarities.



Discussion

In this paper we have investigated the possibility to study non-invariance as a continuous indicator, rather than as a dichotomy. We have shown that using continuous invariance indicators can both be used to diagnose cross-cultural invariance properties of scales under study and identify possible reasons for non-invariance, but also can themselves provide a meaningful source of data for cross-cultural research, echoing the points raised by other researchers (e.g. Medina et al., 2009).

The results of the current analysis show, similar to previous studies that reverse-coded items, problematic measurement-behaviors across cultures (Croucher & Kelly, 2019; Hult et al., 2008; Karl et al., 2020). Our current analysis goes beyond these previous findings by showing that the magnitude of non-invariance between countries can be predicted from systematic features. In the current study we used the Jaeger PMI index (2019), which captures the linguistic difference between a country's main language from other languages. The finding that non-invariance between countries increases systematically with linguistic distance has several potential implications.

First, while the subjective happiness scale showed on average only relatively small differences between countries, a metric invariance test suggested that the structure was not identical. Importantly, linguistic effects were detectable when correlating these effect size estimates of loading variations with linguistic distance metrics. This raises the question if the strength of linguistic effects increases with constructs that show greater cross-cultural differences. Second, this finding provides further credence to claims that cross-cultural differences in measurement properties are not random and with increasing linguistic and cultural distance meaningful comparisons become more difficult to achieve (Boer et al., 2018; Fischer & Poortinga, 2018). Third, the effect of linguistic distance was more pronounced in our data for negatively worded items, supporting previous observations that negations show particular problems for invariance tests. Using this linguistic indicator, we demonstrated that linguistic similarity indeed plays a larger role for negatively, compared to positively, phrased items. This finding implies that negatively worded items might be appropriate in cultural comparisons of samples with low linguistic distance, but researchers might consider omitting negatively scored items with increasing linguistic distance, given that negatively worded items tend to challenge even configural invariance (Zhang et al., 2020). These patterns highlight the need for cross-cultural researchers to engage more deeply with linguistic differences between cultures beyond translation (for an example see: Hodel et al., 2017) and to identify potential features that are especially susceptible to linguistic effects (for example the use of double negation, Déprez et al., 2015). These analyses also raise questions about the nature of psychological constructs and their dependence on linguistic representations.

Beyond probing the comparability of items across countries, we also demonstrated how average non-invariance between pairwise countries can be used to meaningfully cluster countries. We found that both GDP and average temperature can be used to explain country clustering, which supports previous arguments about the importance of both wealth and

climate for cultural differences in general. Our data shows that these features also influence the relative invariance parameters across samples. The linguistic distance was informative for predicting pairwise differences, particularly for negatively worded items. However, it was less informative for clustering country level data when using average invariance parameters. This raises the intriguing possibility that it is possible to capture information on item bias both at the item level as well as information on generalized item bias at the instrument or survey level across a wider range of measures for countries, allowing the systematic capture of differential item use at both levels. Because pairwise effect sizes can be estimated for individual items, more specific hypotheses at the item level can be formulated in the future, allowing for the investigation of cultural and environmental effects on specific sets of items as well as focusing on cultural biases more broadly at instrument level, as has been done in previous research.

Overall, we hope that the current paper helps to challenge the commonly held conception that non-invariance is an unnecessary barrier for cross-cultural research. Instead, we propose that cross-cultural psychologists should engage more deeply and systematically with non-invariance. We believe that the ongoing development of continuous non-invariance indicators based on effect size measures allows for the formulation of predictive theories that provide explanatory mechanisms for cross-cultural differences in the use of psychometric scales.

References

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

Brandt, M. J., He, J., & Bender, M. (2021). Registered Report: Testing Ideological Asymmetries in Measurement Invariance. *Assessment*, *28*(3), 687–708. https://doi.org/10.1177/1073191120983891

Byrne, B. M., & Matsumoto, D. (2021). The Evolution of Multigroup Comparison Testing across Culture: Past, Present, and Future Perspectives. In B. G. Adams & M. Bender (Eds.), *Methods and Assessment in Culture and Psychology* (pp. 296–318). Cambridge University Press. https://doi.org/10.1017/9781108675475.015

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Croucher, S. M., & Kelly, S. (2019). Measurement in Intercultural and Cross-Cultural Communication. In *Communication Research Measures III*. Routledge. https://doi.org/10.4324/9780203730188-10

Deaton, A. (2008). Income, Health, and Well-Being around the World: Evidence from the Gallup World Poll. *Journal of Economic Perspectives*, *22*(2), 53–72. https://doi.org/10.1257/jep.22.2.53

Déprez, V., Tubau, S., Cheylus, A., & Espinal, M. T. (2015). Double Negation in a Negative Concord language: An experimental investigation. *Lingua*, *163*, 75–107. https://doi.org/10.1016/j.lingua.2015.05.012

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, *10*, 1507. https://doi.org/10.3389/fpsyg.2019.01507

Fischer, R., Karl, J. A., Fontaine, J. R. J., & Poortinga, Y. H. (2022). Evidence of Validity Does not Rule out Systematic Bias: A Commentary on Nomological Noise and Cross-Cultural Invariance. *Sociological Methods & Research*, 00491241221091756. https://doi.org/10.1177/00491241221091756

Fischer, R., & Poortinga, Y. H. (2018). Addressing methodological challenges in culture-comparative research. *Journal of Cross-Cultural Psychology*, *49*(5), 691–712. https://doi.org/10.1177/0022022117738086

Fischer, R., & Smith, P. B. (2021). How Far Can Measurement Be Culture-Free? In B. G. Adams & M. Bender (Eds.), *Methods and Assessment in Culture and Psychology* (pp. 319–340). Cambridge University Press. https://doi.org/10.1017/9781108675475.016

Fischer, R., & Van de Vliert, E. (2011). Does Climate Undermine Subjective Well-Being? A 58-Nation Study. *Personality and Social Psychology Bulletin*, *37*(8), 1031–1041. https://doi.org/10.1177/0146167211407075

Fontaine, J. (2005). Equivalence. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 1, pp. 803–813). https://doi.org/10.1016/B0-12-369398-5/00116-X

Gardiner, G., Lee, D., Baranski, E., & Funder, D. (2020). Happiness around the world: A combined etic-emic approach across 63 countries. *PLoS ONE*, *15*(12), e0242718. https://doi.org/10.1371/journal.pone.0242718

Georgas, J., van de Vijver, F. J. R., & Berry, J. W. (2004). The Ecocultural Framework, Ecosocial Indices, and Psychological Variables in Cross-Cultural Research. *Journal of Cross-Cultural Psychology*, *35*(1), 74–96. https://doi.org/10.1177/0022022103260459

Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2019). Evaluation of Six Effect Size Measures of Measurement Non-Invariance for Continuous Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–12. https://doi.org/10.1080/10705511.2019.1689507

Hodel, L., Formanowicz, M., Sczesny, S., Valdrová, J., & von Stockhausen, L. (2017). Gender-Fair Language in Job Advertisements: A Cross-Linguistic and Cross-Cultural Analysis. *Journal of Cross-Cultural Psychology*, *48*(3), 384–401. https://doi.org/10.1177/0022022116688085

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values.* Sage.

Hult, G. T. M., Ketchen, D. J., Griffith, D. A., Finnegan, C. A., Gonzalez-Padron, T., Harmancioglu, N., Huang, Y., Talay, M. B., & Cavusgil, S. T. (2008). Data equivalence in cross-cultural international business research: Assessment and guidelines. *Journal of International Business Studies*, *39*(6), 1027–1044. https://doi.org/10.1057/palgrave.jibs.8400396

Inglehart, R. (1997). *Modernization and Postmodernization*. Princeton University Press.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion semantics show both cultural variation and universal structure. *Science*. https://doi.org/10.1126/science.aaw8160

Karl, J. A., Méndez Prado, S. M., Gračanin, A., Verhaeghen, P., Ramos, A., Mandal, S. P., Michalak, J., Zhang, C.-Q., Schmidt, C., Tran, U. S., Druica, E., Solem, S., Astani, A., Liu, X., Luciano, J. V., Tkalčić, M., Lilja, J. L., Dundas, I., Wong, S. Y. S. Y., … Fischer, R. (2020). The cross-cultural validity of the Five-Facet Mindfulness Questionnaire across 16 countries. *Mindfulness*, *11*, 1226–1237. https://doi.org/10.1007/s12671-020-01333-6

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. SAGE. https://doi.org/10.4135/9781412985130

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 231–248. https://doi.org/10.1348/000711003770480020

Matsumoto, D., Grissom, R. J., & Dinnel, D. L. (2001). Do Between-Culture Differences Really Mean that People are Different?: A Look at Some Measures of Cultural Effect Size. *Journal of Cross-Cultural Psychology*, *32*(4), 478–490. https://doi.org/10.1177/0022022101032004007

Medina, T. R., Smith, S. N., & Long, J. S. (2009). Measurement Models Matter: Implicit Assumptions and Cross-national Research. *International Journal of Public Opinion Research*, *21*(3), 333–361. https://doi.org/10.1093/ijpor/edp037

Messick, S. (1991). Psychology and methodology of response styles. In *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 161–200). Lawrence Erlbaum Associates, Inc.

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

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. https://doi.org/10.1037/a0022955

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best Practices in Exploratory Factor Analysis. In J. Osborne, *Best Practices in Quantitative Methods* (pp. 86–99). SAGE Publications, Inc. https://doi.org/10.4135/9781412995627.d8

R Core Team. (2018). R: A Language and Environment for Statistical Computing.

Rosseel, Y. (2012). {lavaan}: An {R} package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222. https://doi.org/10.1016/j.hrmr.2008.03.003

Shi, D., Song, H., & Lewis, M. D. (2019). The Impact of Partial Factorial Invariance on Cross-Group Comparisons. *Assessment*, *26*(7), 1217–1233. https://doi.org/10.1177/1073191117711020

Stanca, L. (2010). The Geography of Economics and Happiness: Spatial Patterns in the Effects of Economic Conditions on Well-Being. *Social Indicators Research*, *99*(1), 115–133. https://doi.org/10.1007/s11205-009-9571-1

Travel Weather Averages (Weatherbase). (2022). Weatherbase. http://www.weatherbase.com/

Tsai, J., & Park, B. (2014). The cultural shaping of happiness: The role of ideal affect. In *Positive emotion: Integrating the light sides and dark sides* (pp. 345–362). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199926725.003.0019

Uchida, Y., Norasakkunkit, V., & Kitayama, S. (2004). Cultural constructions of happiness: Theory and emprical evidence. *Journal of Happiness Studies*, *5*(3), 223–239. https://doi.org/10.1007/s10902-004-8785-9

Van de Vijver, F. J. R. R., & Leung, K. (2021). *Methods and Data Analysis for Cross-Cultural Research* (V. H. Fetvadjiev, J. Fontaine, J. He, & Y. H. Poortinga, Eds.; 2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781107415188

Van de Vliert, E. (2007). Climates Create Cultures. *Social and Personality Psychology Compass*, *1*(1), 53–67. https://doi.org/10.1111/j.1751-9004.2007.00003.x

Van de Vliert, E., & Van Lange, P. A. (2020). Latitudinal gradients as scientific tools for psychologists. *Current Opinion in Psychology*, *32*, 43–46. https://doi.org/10.1016/j.copsyc.2019.06.018

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Welzel, C., Brunkert, L., Kruse, S., & Inglehart, R. F. (2021). Non-invariance? An Overstated Problem With Misconceived Causes. *Sociological Methods & Research*, 0049124121995521. https://doi.org/10.1177/0049124121995521

Zager Kocjan, G., Jose, P. E., Sočan, G., & Avsec, A. (2021). Measurement Invariance of the Subjective Happiness Scale Across Countries, Gender, Age, and Time. *Assessment*, 1073191121993558. https://doi.org/10.1177/1073191121993558

Zhang, B., Luo, J., Chen, Y., Roberts, B., & Drasgow, F. (2020). *The road less traveled: A cross-cultural study of the negative wording factor in multidimensional scales*. PsyArXiv. https://doi.org/10.31234/osf.io/2psyq