Journal of Quantitative Criminology
https://doi.org/10.1007/s10940-023-09578-9

ORIGINAL PAPER

Can We Compare Attitudes Towards Crime Around the World? Assessing Measurement Invariance of the Morally Debatable Behavior Scale Across 44 Countries

Sandy Schumann[1] · **Michael Wolfowicz**[1]

Accepted: 10 August 2023
© The Author(s) 2023

Abstract

Objectives We aim to encourage scholars who conduct cross-national criminological studies to routinely assess measurement invariance (MI), that is, verify if multi-item instruments that capture latent constructs are conceptualized and understood similarily across different populations. To promote the adoption of MI tests, we present an analytical protocol, including an annotated R script and output file. We implement the protocol and, doing so, document the first test of configural, metric, and scalar invariance of the three-factor Morally Debatable Behavior Scale (MDBS).

Methods We worked with data from wave seven of the World Values Survey (WVS). Applying multi-group confirmatory factor analyses, we, first, explored invariance of the MDBS in 44 countries ($N = 59{,}482$). Next, we conducted analyses separately for seven South-american, six South-east Asian, six East-asian, two North American and Australasian, and all four Anglophone countries.

Results The MDBS displays an overall lack of invariance. However, we confirmed configural invariance of the MDBS for the South-east Asian sample, metric invariance in the sample of Anglophone countries, and scalar invariance for the Australasian and North American countries.

Conclusions Wave seven of the WVS can be used for latent mean score comparisons of the MDBS between the Australasian and North American countries. Associative relationships can be compared in the larger Anglophone sub-sample. Taken together, MI must be tested, and cannot be assumed, even when analyzing data from countries for which previous research has established cultural similarities. Our protocol and practical recommendations guide researchers in this process.

Keywords Measurement invariance · Morally debatable behavior scale · Cross-national research · Comparative criminology

✉ Sandy Schumann
s.schumann@ucl.ac.uk

Michael Wolfowicz
michael.wolfowicz@mail.huji.ac.il

[1] Department of Security and Crime Science, University College London, 35 Tavistock Square, London WC1H 9EZ, UK

Published online: 04 September 2023

Introduction

Following calls for the internationalization of criminology, cross-national studies,[1] in particular such that draw on large-*n* surveys, have become increasingly common (Barberet 2007; Messner 2021). This is a welcomed development, given the substantial benefits of those analyses. First, cross-national research allows for the testing of ostensibly general theories of crime that should apply in different contexts; relevant country-level moderators of the causes of crime can be identified when a wide variety of settings is considered (Bennett 1980; Nivette 2021; Messner 2015). Relatedly, studying crime across diverse populations could highlight existing biases, pointing perhaps to unjustified generalizations and assumptions of uniformity (Aebi and Linde 2015). In turn, if cross-national research shows similar dynamics of, for instance, crime trends in several countries, specific national explanations must be abandoned in favor of universal theories. That is, based on the results of cross-national studies, new theoretical frameworks can be established (Bennett 2004; 2009; Karstedt 2001). Additionally, investigating the impact of interventions to prevent and reduce crime in various countries provides insights on whether what constitutes best practices in one setting should be applied in other places as well (Tonry 2015). Opportunities for crucial collaborations between law enforcement and criminal justice institutions can then be identified to address transnational crime (Bennett 2004).

Although it offers significant promise, cross-national research is not without (methodological) challenges. This paper focuses on a particular area of concern that affects cross-national (survey) studies that employ multi-item scales to assess one or more underlying latent constructs. Multi-item scales are often preferred over single item measures as the latter are more likely to exhibit a larger measurement error and lower predictive validity (Diamantopoulos et al. 2012). Having said this, the applicability of multi-item scales in cross-national research is restricted if the instruments are not conceptualized and understood in the same manner in different populations, that is, if measurement invariance (MI) of the scale(s) is not achieved (Davidov et al. 2014; Van de Schoot et al. 2012; Vandenberg and Lance 2000). Notably, different levels of MI—explained in more detail below—are required to establish accurate estimates of country-differences of latent (i.e., scalar invariance) or manifest (i.e., uniqueness) mean scores or regression coefficients (i.e., metric invariance) (Chen 2008; Schmitt et al. 2011).

Methods for assessing MI have been advanced considerably in the last five decades (Leitgöb et al. 2022; van de Schoot et al. 2012). Additionally, MI tests have been implemented in different disciplines to explore the equivalence of various scales across, for instance, countries, gender, age and racial groups (e.g., Bieda et al. 2017 (happiness); Dong and Dumas 2020 (personality measures); Wicherts et al. 2005 (test performance)). Criminologists, specifically, have examined MI, among others, for instruments that capture fear of crime (Pauwels and Pleysier 2005; Pleysier et al. 2004), collective efficacy (Gerstner et al. 2019), or self-control (Pechorro et al. 2022). Reviewing the latter work, it is, however, evident that the tests of measurement invariance were not always adequate or relied on inconsistent standards to judge whether certain levels of equivalence were attained (see also Schmitt and Kuljanin 2008; Vandenberg and Lance 2000). Given the importance of measurement equivalence for the accuracy of the results of many cross-national

[1] We refer to cross-national studies as such that include data from two or more countries, thus, combining a cross- and multi-national scope (Bennett 2004).

criminological studies, it, therefore, appears important to provide accessible guidance on how to assess MI.

This paper presents such a practical primer, combining insights from more technical or less detailed reviews (see Fischer and Karl 2019; van de Schoot et al. 2012) in a 'step-by-step' protocol that is accompanied by an annotated R code script and output file. More precisely, we introduce the most commonly used approach for testing measurement invariance—multi-group confirmatory factor analysis, which relies on a pre-determined factor structure of a scale. To illustrate the application of the analytical protocol, we investigate invariance of the three-factor Morally Debatable Behavior Scale (MDBS; Harding and Phillips 1986), administered in 44 countries in wave seven of the World Values Survey (WVS; Haerpfer et al. 2022). Two previous studies have explored MI of a two-factor and three-factor version of the MDBS. However, one study, using a global sample, only tested lower (i.e., metric) levels of invariance (Vauclair and Fischer 2011) such that the accuracy of comparisons of latent scores cannot be concluded. The second study, relying on European data, did not present details about the type of equivalence that was attained (Moors and Wennekers 2003). We advance this work and assess configural, metric, and scalar equivalence of the MDBS in 44 countries. Furthermore, we aim to demonstrate that invariance of a scale should never be assumed but must be tested even for populations who reside in close proximity and share historical ties or for which previous research has determined cultural similarities. To emphasize this point, we also examine measurement invariance of the MDBS separately for countries from distinct geographic and linguistic regions (i.e., countries in South America, East Asia, South-East Asia, North America, Australasia, and Anglophone countries).

Criminological Research Using Cross-National Survey Data

Cross-national criminological studies seek to describe and explain variations in crime, its antecedents, and related phenomena, such as victimization rates or fear of crime, around the world (*descriptive* and *analytic* approach; Bennett 2004). Initially restricted by limited access to relevant samples, progress has been made regarding the scope and quality of data that is employed (see LaFree 2021). Notably, several secondary datasets of rigorous large-*n* surveys, including populations that were previously under-represented, are now publicly available for cross-national criminological research (e.g., International Crime Victims Survey (Van Dijk et al. 1990), the International Self-report Delinquency Surveys (ISRD; Marshall et al. 2022), the Demographic and Health Surveys (covering gender/domestic violence; The DHS Program 2023), the World Values Survey; see Nivette 2021).

Focusing specifically on the World Values Survey, which is described in more detail below, a broad range of research questions have been explored. Chon (2021) documented, for instance, individual- and country-level predictors of attitudes towards gender-based violence across 37 countries (see also Herrero et al. 2017; Tausch 2019; Thulin et al. 2021). Including samples from 48 and 31 countries respectively, less favorable attitudes towards democracy and lower levels of self-efficacy were found to predict stronger justification of terrorism and politically motivated violence (Martinez et al., 2022; Julkif 2022). Data from the World Values Survey has also been used to test institutional-anomie theory in up to 74 countries (e.g., Hirtenlehner et al. 2013; Rogers and Pridemore 2022; Zito 2019). Combining survey data with police-recorded offending rates from 55 nations, there is further

evidence that attitudes towards violence, as measured in the WVS, shape the relationship between firearm prevalence and firearm related homicides (Kovandzic and Kleck 2022).

Most researchers readily admit that the data of the aforementioned work are, albeit broad in geographical scope, not without limitations. A common challenge for cross-national surveys are inconsistent sampling and data collection procedures that jeopardize the comparability of samples and results (Davidov et al. 2014). However, as Rodriguez and colleagues (2015) highlighted, even if those methods are standardized, it also must be ensured that the chosen instrument(s)—the scales and questions—can 'travel' or, put simply, are understood in the same way by different populations. This matter does not refer to the need for rigorous (back) translation of questions, which is, of course, relevant as well. Instead, the authors emphasize that the extent to which an activity has a predominantly illegal connotation might vary between settings. For example, in Venezuela, other than in Western countries, 'painting on a wall (graffiti)' is often part of a normative political or community activity (i.e., painting graffiti is an etic concept; Triandis 1978) (Rodriguez et al., 2015). If 'graffiti' was to be included as an item in a scale that seeks to capture an underlying latent concept of 'delinquency', the measure would not be comparable between Venezuela and Western countries. With the latter point, and although they did not explicitly reference the term, Rodriguez and colleagues (2015) raise an issue known as *measurement invariance*.

Measurement Invariance

Measurement invariance (or equivalence) implies that a multi-item instrument, comprised of one or more latent factors, "evokes the same conceptual frame of reference" (Vandenberg and Lance 2000, p. 9) or is interpreted and responded to in the same way across different populations (or data collection points[2]) (Byrne and Watkins 2003; Putnick and Bornstein 2016; Van de Schoot et al. 2012). Specifically, if respondents from different populations exhibit, for instance, the same level of endorsement of an attitude or behavioral intention, they should complete the instrument in the same way (Davidov et al. 2014). Four types of measurement invariance are distinguished – configural, metric, and scalar invariance, as well as invariance of uniqueness (Schmitt et al. 2011). Depending on the type of MI that is attained, different analytical procedures can be confidently conducted with the respective data. Below, we introduce the four levels of MI following the generalized structural equation modelling approach where a latent construct[3] η is defined by an intercept α as well as n observed/manifest variables $x_1,\ldots, x_n$ and an error term ε (Eq. 1) (Jöreskog 1971). The measurement relationships between the manifest and latent variables are specified by the factor loadings $\beta_{1,\ldots,}\beta_n$ (Eq. 1).

$$\eta = \alpha + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \tag{1}$$

Configural (or nonmetric; Widaman and Reise 1997) invariance suggests that the basic organization of an instrument (i.e., the number of latent factors and the number of

[2] Although we focus on measurement invariance in the context of cross-national research, MI is equally relevant for single-country studies that rely on several waves, and the repeated use of the same measurement tool. Over time, measures may be interpreted differently; therefore, equivalence over time should be examined.

[3] This example focuses on one latent variable. The same principles apply when more than one latent construct is proposed.

loadings—both significant and non-significant— from particular manifest variables on each latent factor) is equivalent in all sub-groups M, here indexed by j (Rutkowski and Svetina 2014; Widaman and Reise 1997) (Eq. 2).

$$\eta^j = \alpha^j + \beta_1^j x_1^j + \cdots + \beta_n^j x_n^j + \varepsilon^j, \text{ for } j = 1 \text{ to } M \tag{2}$$

For example, if a measure specifies one latent variable 'delinquency' that is thought to be defined by three manifest variables that capture the frequency of certain activities, including 'graffiti', configural invariance indicates that in the respective sub-groups, the three tested behaviors reflect only one underlying construct, and that spraying graffiti is considered an indicator of delinquency everywhere.

Metric (or weak factorial; Meredith 1993) equivalence stipulates that beyond the requirement of configural invariance, the precise numeric values of the factor loadings, that is, the strength of the relationships between all manifest and the latent variable(s), are *also* the same in all sub-samples (Eq. 3; Rutkowski and Svetina 2014).

$$\beta_n^j = \beta_n^k, \text{ for all } n \text{ and } j, k = 1, \ldots, M \tag{3}$$

In the case of the previously introduced example, metric invariance would imply that across all sub-groups, respondents have a similar understanding of whether (i.e., configural invariance) *and* the extent to or exact strength with which (i.e., metric invariance) the frequency of spraying graffiti and two other activities reflect one underlying construct of delinquency. Confirming metric invariance indicates that "the unit of measurement is equal across groups" (Schmitt et al. 2011, p. 413); a one-unit change on a scale implies the same meaning in all sub-groups. Consequently, the variance of the latent construct(s) as well as relationships with other variables (i.e., regression or correlation coefficients) are expected to be estimated in the same way in all groups (Guenole and Brown 2014), and are, thus, comparable (Meuleman 2012). Metric invariance also allows for testing different types of regression models (i.e., mixed/multi-level, hierarchal models) in aggregated multi-country datasets (Davidov et al. 2014). Conversely, simulation studies have found that when metric noninvariant measures are selected, estimates of regression coefficients are inaccurate (Chen 2008; Guenole and Brown 2014). Chen's (2008) analysis, comparing an American and Chinese sub-sample, for example, showed that a metric noninvariant predictor resulted in between-group differences in standardized regression coefficients of up to 0.30[4]; slopes were underestimated for the American and overestimated for the Chinese sub-sample.

Scalar (or strong factorial) invariance introduces further restrictions to the metric model. Notably, the intercept, or constant, of the measurement model is expected to be the same in all sub-groups (Eq. 4).

$$\alpha^j = \alpha^k \text{ for all } j, k = 1, \ldots, M \tag{4}$$

To understand the implications of and the need for scalar invariance better, it is useful to recall the meaning of the intercept in a regression model: it is the value of the dependent variable—here, the latent variable—when all predictors—here, the manifest items—are zero. Returning to the previous example, the equivalence of intercepts in different groups

[4] This value refers to a scenario where 87.5% of factor loadings were noninvariant and sample sizes differed; given equal sample sizes and only 25% if invariant factor loadings, slope differences were negligible at .06.

implies that the mean level of the construct delinquency is the same across sub-groups when the three measured activities, including graffiti, are all recorded with zero frequency, $x_1—x_n = 0$. In other words, the point of origin of the measure of the latent construct delinquency is the same across sub-samples. Consequently, any observed between-group differences in delinquency can be explained by differences in the frequency with which the three measured activities are performed (Davidov et al. 2014). Studies that aim to compare means of latent variables across countries or contexts must therefore attain scalar invariance; between-group differences might otherwise be overestimated (Steinmetz 2013; Vandenberg and Lance 2000). Specifically, under conditions of metric or scalar invariance, mean scores were found to be artificially elevated for groups that attained higher factor loadings causing spurious group differences (Chen 2008). Those discrepancies in estimates are exacerbated when sample sizes differ between study units (Chen 2008).

Lastly, for accurate comparisons of observed group means, that is, means of manifest scores, *residual* (or strict or invariant uniqueness) equivalence is also required. In addition to the conditions of scalar equivalence, error variances and the variances of items that are not shared with the latent variable are expected to be the same across all countries (Eq. 5; Widaman and Reise 1997).

$$\varepsilon^j = \varepsilon^k \text{ for all } j, k = 1, \ldots, M \tag{5}$$

In practice, invariant uniqueness is rarely assessed empirically, which might be explained by the fact that most follow-up analyses compare latent means or apply regression models.

The Present Study

Over the last five decades, the implementation of MI tests has been advanced; most common statistical programs (e.g., Mplus, STATA, R) offer options to examine MI (Leitgöb et al. 2022; van de Schoot et al. 2012; Vandenberg and Lance 2000; Vandenberg and Lance 2000). Similar to developments in other disciplines (e.g., Bieda et al. 2017; Dong and Dumas 2020; Wicherts et al. 2005), measurement invariance has been assessed in several criminological studies. Spencer et al. (2005) demonstrated that the factor structure of the Behavioral Problem Index was not the same for different ethnic groups in the US. The authors identified the source of this variance, that is, the noninvariant items, and created new sub-scales for which between-group comparisons were feasible. This study illustrates that measurement invariance tests are not only relevant for cross-national research but also for work that aims to compare social groups within one context. Relatedly, the factor structure of the low self-control scale was explored for female and male participants in a Portuguese sample (Pechorro et al. 2022). Moreover, equivalence of a scale that captures fear of crime was investigated for different linguistic groups, as well as female and male participants, in Belgium (Pleysier et al. 2004). Gerstner et al. (2019) further examined equivalence of a measure of collective efficacy, introduced originally in the Chicago Study (Sampson et al. 1997), in Germany and Australia. They demonstrated only metric invariance and concluded that latent means of the measure should not be compared across the two settings. Subsequent analyses were conducted separately for each sample. Similarly, Nivette et al. (2020) documented that a measure of legal cynicism attained metric invariance in a sample of adolescents from Brazil, Uruguay, and Switzerland. Consequently, the

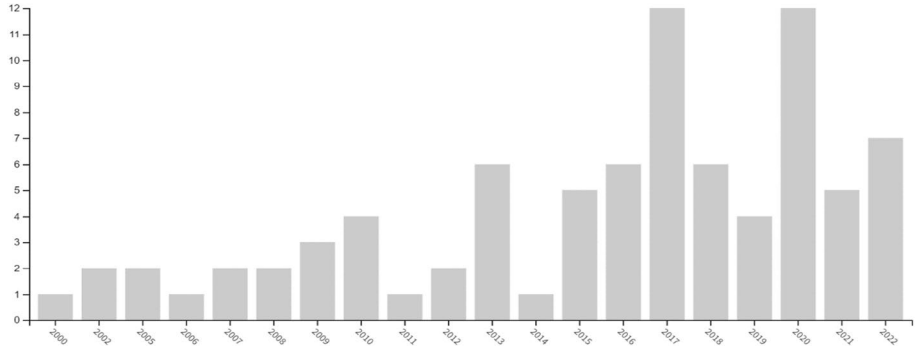

Fig. 1 Trends of publications that refer to 'measurement invariance' between 2000 and 2022. *Source*: Web of Science, search terms: ((ALL = (crime)) OR (ALL = (offending)) OR (ALL = (delinquency))) AND (ALL = (measurement invariance))

authors did not compare latent means across all three settings but were able to assess the relative importance of different antecedents in the three countries.

The aforementioned studies are promising in that they suggest that MI tests are not ignored in criminology. Having said this, as shown in Fig. 1, measurement invariance has received overall limited attention in the field. Further, as also highlighted in previous review articles (focused on organizational research), the implementation of measurement invariance tests is not always adequate and varies widely (Schmitt and Kuljanin 2008; Vandenberg and Lance 2000). Spencer and colleagues (2005), for instance, compared a model that introduced restrictions in line with scalar invariance with a model where factor loadings and intercepts were allowed to vary between ethnic groups. They then assessed the $\chi 2$ scores of the two models. As will be described below, this approach is not suited to identify invariance. Pechorro and colleagues (2022) simply tested the fit of models with different numbers of latent factors in female and male sub-samples. At best, this could suggest configural invariance. Pleysier and colleagues (2004) considered multi-group confirmatory factor analysis but reviewed only the root mean square error of approximation (RMSEA) (using a threshold of < 0.05) as a fit statistic. Nivette and colleagues also relied on the RMSEA using < 0.06 as a suitable threshold to judge each model's fit; they further evaluated changes in the comparative fit index (CFI; Bentler 1990) to compare models with different levels of restrictions.

We cannot determine the reasons for these inconsistencies in MI tests. One explanation could be that existing guidance for the assessment of measurement invariance (e.g., Fischer and Karl 2019; Leitgöb et al. 2022; van de Schoot et al. 2012) might seem overly technical, especially for scholars who have not yet worked with the respective software or have never applied structural equation modelling techniques. We believe that more researchers who conduct cross-national criminological studies would implement MI tests (accurately) if they were able to rely on accessible recommendations.

One aim of this paper is to provide such a primer. More precisely, we seek to demonstrate how measurement invariance can be examined by presenting the most commonly used approach, multi-group confirmatory factor analysis, as a 'step-by-step' protocol and by providing the script required to conduct the analysis in R. We illustrate the implementation of the protocol with a concrete example of an instrument for which the underlying

(latent) factor structure of the scale has been established,[5] specifically, the multi-item Morally Debatable Behavior Scale (MDBS) as fielded in 44 countries in wave seven in the World Values Survey. The MDBS assesses the extent to which individuals justify actions that are morally contestable or illegal with Likert-type answer options. The original scale, developed in Western Europe in the 1980s (Harding and Phillips 1986), was revised in the mid-1990s for an American sample (Katz et al. 1994; MDBS-Revised (MDBS-R)). Initially devised as a two-factor scale, the MDBS-R was found to load on three factors: honesty-dishonesty (e.g., claiming welfare benefits you're not entitled to; cheating on your taxes if you have a chance), personal-sexual (e.g., divorce; prostitution), and legal-punitive (e.g., killing in self-defense, capital punishment).[6] The honesty-dishonesty sub-scale refers to attitudes towards behavior that violates social norms and that is sanctioned by the state; the personal-sexual factor reflects the judgement of behavior in the realm of personal freedom that might be contested by social norms but is not illegal (Marozzi 2021; Vauclair and Fisher 2011). The legal-punitive scale of the MDBS-R included originally actions of physical violence that are permissible and legal. However, in more recent adaptions, like the one studied in the present research, the third sub-scale examines the justification of illegal, interpersonal and group-based violence.

To date, the MDBS has been widely used in cross-national research to explore topics such as fraud morality, support for violent extremism, or attitudes towards different types of interpersonal violence (e.g., Chon 2021; Julkif 2022; Martinez et al., 2022).[7] However, to our knowledge, only two studies have explored MI of the MDBS-R. Vauclair and Fischer (2011) showed metric invariance for a 10-item, two-factor, version of the scale in 56 countries. The authors did not go on to test scalar invariance such that it cannot be confirmed if cross-national comparisons of the latent scale scores would have been warranted. Moors and Wennekers (2003), using data from the European Values Studies in 12 Western, Northern, and Southern European countries and a nine-item, three-factor, MDBS did indeed conclude that such between-group comparisons were not justified. However, given the analytical approach that was chosen, conclusions about configural or metric invariance were not determined.

To address those gaps in the literature, the present research investigates whether configural, metric, and scalar measurement invariance of the three-factor version of the MDBS can be attained in a sample of 44 countries in which the instrument was fielded in wave seven of the WVS (*Research Question 1*). It could of course be argued that expectations of configural, metric, and scalar MI are unrealistic in such a highly diverse sample (Sokolov 2021), and it would perhaps not be surprising that respondents from countries as Greece, Pakistan, and Thailand conceptualize attitudes towards group-based violence or behavior that is sanctioned by the state in different ways. In fact, the legality of several items on the MDBS (e.g., prostitution, abortion) varies between countries. Indeed, researchers might intuitively expect that invariance of a scale is more reasonable in sub-samples that are in close geographic proximity and have historical ties or for which previous research has

[5] See Fischer and Karl (2019) for an overview of MI assessment in the context of exploratory measurement models.

[6] See Halpern (2001), who showed three factors labeled self-interested morality, legal-illegal, and personal-sexuality.

[7] Further applications of the MDBS in previous WVS waves include work on associations between trust, generalized morality, and economic performance (Franke and Nadler 2008; James 2015), civic morality and its predictors (Letki 2006), or religiosity and moral values (Storm 2016), as well as questions of moral universalism vs relativism (Vauclair and Fisher 2011).

determined cultural similarities. Our aim is to demonstrate that even in those scenarios MI must be investigated and cannot be taken for granted. Hence, we explore if configural, metric, and scalar invariance of the MDBS is supported in five regional sub-samples: countries in South America, South-East Asia, East Asia, North America, as well as countries in Australasia (*Research Question 2a*). Additionally, we consider a sub-sample that includes both the North American and Australasian countries (i.e., Canada, the United States, Australia, and New Zealand) (*Research Question 2b*). These four countries have been described as being similar on several cultural dimensions (e.g., indulgence, uncertainty avoidance, individualism, and masculinity; Hofstede 2022), as well as with respect to certain values (e.g., harmony, hierarchy, and egalitarianism; Schwartz 1992).

Method

Data and Sample

We drew on data from wave 7 of the World Values Survey (WVS), which was fielded from 2017 to 2021 (Haerpfer et al. 2020). Depending on the country, full probability samples (above 18 years old) or nationally representative random samples based on multi-stage territorial stratified selection were recruited. Data were collected primarily through in-person interviews. The entire dataset, available at the time of analysis, includes responses from $N = 74{,}301$ individuals from 51 countries, with an average of $N = 1200$ respondents per country (Table 1). For the purposes of the current study, it was necessary to remove observations which had any missing or incomplete responses on the MDBS (for detail see Analytical Procedure), resulting in an analytical sample of $N = 59{,}482$ respondents from 44 countries.

It has been recommended that MI tests rely on data that was collected using largely the same method in all sub-groups (Leitgöb et al. 2022). Furthermore, a sub-group sample size larger than $n = 200$ is suggested (Fischer and Karl 2019). Our dataset complies with those recommendations.

Measure

The analysis was based on an adaption of the Morally Debatable Behaviors Scale-Revised. In wave seven, the WVS fielded 10 items that were included in the MDBS-R (Katz et al. 1994) and nine additional items that were wave-specific (Table 2). The items can be grouped into three sub-scales, namely, (dis)honesty (i.e., non-violent illegal behaviors), (il) legal behaviors (i.e., violent illegal behaviors), and personal-sexual related behaviors (e.g., Halpern 2001) (Table 2). Each item was rated on a 10-point Likert-type scale, indicating to what extent the respective behaviors were considered justifiable (1 = *Never justifiable*, 10 = *Always justifiable*). Importantly, the answer options for the scale were the same in all countries; were this not the case, MI could not be tested (Leitgöb et al. 2022).

In line with previous studies, we treated the items of the MDBS as interval-scaled (e.g., Franke and Nadler 2008; James 2015; Letki 2006; Marozzi 2021; Vauclair and Fisher 2011; Storm 2016). We acknowledge that some may consider the items as ordinal-scaled. Additionally, the items have also been dichotomized (Chon 2021; Julkif 2022; Martinez et al., 2022), given that the distribution of the measure is often skewed. The overall procedure of assessing measurement invariance as described below, using the R script that

Table 1 Sample overview

Country	*N* full survey (*N* excluding missing values on MDBS)	Country	*N* full survey (*N* excluding missing values on MDBS)
Andorra	1004 (946)	Macao	1023 (987)
Argentina	1003 (717)	Malaysia	1313 (1313)
Australia	1813 (1638)	Mexico	1739 (1590)
Bangladesh	1200 (1200)	New Zealand	1057 (751)
Bolivia	2067 (1441)	Nicaragua	1200 (1200)
Brazil	1762 (1259)	Nigeria	1237 (1113)
Myanmar	1200 (1191)	Pakistan	1995 (1762)
Canada	4018 (4018)	Peru	1400 (1170)
Chile	1000 (794)	Philippines	1200 (1189)
China	3036 (2931)	Puerto Rico	1127 (1053)
Taiwan	1223 (1222)	Romania	1257 (1036)
Colombia	1520 (1520)	Russia	1810 (1358)
Cyprus	1000 (395)	Serbia	1046 (925)
Ecuador	1200 (1008)	Singapore	2012 (1924)
Ethiopia	1230 (1051)	Vietnam	1200 (1200)
Germany	1528 (1313)	Zimbabwe	1215 (1179)
Greece	1200 (1007)	*Tajikistan*	*1200 (0)*
Guatemala	1203 (1058)	Thailand	1500 (1422)
Hong Kong	2075 (2010)	Tunisia	1208 (1178)
Indonesia	3200 (3107)	*Turkey*	*2415 (0)*
Iran	*1499 (0)*	Ukraine	1289 (632)
Iraq	*1200 (0)*	*Egypt*	*1200 (0)*
Japan	1353 (1031)	United States	2596 (2443)
Kazakhstan	1276 (917)		
Jordan	*1203 (0)*		
South Korea	1245 (1245)		
Kyrgyzstan	1200 (1038)		
Lebanon	*1200 (0)*		

Note. Country names in italic refer to countries that are not included in the analytical sample

accompanies the paper, applies regardless of whether the chosen scale is a continuous or categorical measure. However, for ordinal/categorical data the model estimators must be changed; we describe this step in more detail in the next section.

Analytical Procedure

All analyses were conducted using *R* 4.2.1; the package *lavaan* 0.6–11 (Rosseel 2012) was employed for all MI tests. The annotated R code script, the dataset that reflects the analytical sample of our analysis, and an annotated R output file, which allows readers to reproduce our analyses and, by adjusting the code to their project, pursue their own assessments of measurement invariance, are available here: https://tinyurl.com/cua5bbdx. Below, we make reference to particular steps in the R script such that those can be directly linked

Table 2 MDBS-R items in wave seven of the WVS

Item number		Honesty-dishonesty	Personal-sexual	Legal-illegal
1	Claiming government benefits to which you are not entitled	x		
2	Avoiding a fare on public transport	x		
3	Stealing property	*		
4	Cheating on taxes if you have a chance	x		
5	Someone accepting a bribe in the course of their duties	x		
6	Homosexuality		x	
7	Prostitution		x	
8	Abortion		*	
9	Divorce		x	
10	Sex before marriage		*	
11	Suicide		x	
12	Euthanasia		x	
13	Having casual sex		*	*
14	For a man to beat his wife			*
15	Parents beating children			*
16	Violence against other people			*
17	Terrorism as a political, ideological or religious mean			*
18	Political violence			*
19	Death penalty			x

Note. x = items of MDBS-R * = additional items included in wave 7 of the WVS

to the respective procedure in the text. For researchers who have never used R before, we recommend first downloading the software (https://cran.r-project.org/) as well as RStudio (https://posit.co/products/open-source/rstudio/).

We prepared the data by removing all cases with at least one missing value on either MDBS-R item, which included cases with the response 'Don't know', refusal to answers, and cases where the question was not asked/not applicable due to a country-specific filter. Table 1 shows sample sizes before and after missing values were excluded. Where no cases were available for a country (i.e., percentage of missing values equals 100%), one or more items of the MDBS-R were not asked. For instance, 'Prostitution' was not included in the scale in Lebanon; 'Homosexuality' was not part of the instrument in Tajikistan. In Supplementary Material S1 we documented the frequency with which different types of missing data were recorded for each item in each country. Table 1 as well as Table S1.1 highlight a wide variety in the frequency and patterns of missing answers. For instance, in countries like Colombia, Bangladesh, Taiwan, and Canada no type of missing data is present. 'Don't know' was not recorded in Andorra and Australia but in several other countries. Differences in missing data patterns could suggest that certain items are not well understood in particular countries. In the present study, however, we believe that those differences indicate differences in data collection methods within and between countries. Interviewers of the WVS were instructed to not offer the options 'Don't know' or 'refuse' as answer options but recorded the response if the participants voiced it. It is possible that different interviewers within a country varied with respect to how accurately missing answers were interpreted as well as how much they encouraged

participants to report answers after they may have initially not done so. Moreover, it can also be speculated that some countries were keener than others to avoid missing data in the sample and systematically encouraged responding—the fact that several countries report zero missing values would suggest that. Overall, we believe that the missing data patterns are not random. Indeed, Little's test ($\chi 2 = (29{,}132) = 61{,}598.06$, $p < 0.001$) showed that the assumption of missing completely at random is not supported. We also did not find it reasonable to assume a missing at random pattern, which is why list-wise deletion rather than imputation was chosen to deal with the missing data. Implications of this choice are discussed below.

The following analyses were, first, applied in the whole sample and then replicated across the pre-defined sub-samples. MI was examined following a generalized latent variable approach, specifically, employing the widely used multi-group confirmatory factor analysis (MGCFA), where increasing between-group equality restrictions are introduced in several steps (Jöreskog, 1971; Widaman and Reise 1997). As we assumed that continuous observed/manifest variables define, or are indicators to measure, continuous latent constructs, the maximum likelihood estimator was chosen. For analyses that include categorical variables but also apply the protocol and R script presented in this paper, a robust weighted least squares estimator (WLSMV) must be chosen (see Rosseel, 2023 for an implementation in R; see Davidov et al. 2014 and Rhemtulla et al., 2012 for alternative approaches).

In the first step, the fit of the three-factor baseline model was tested in the whole sample (Eq. 6) [R Script: Define and assess fit of baseline measurement model for whole sample].

$$\begin{aligned}
\eta_{Honesty-dishonesty} &= \alpha_{Honesty-dishonesty} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon_{Honesty-dishonesty} \\
\eta_{Personal-sexual} &= \alpha_{Personal-sexual} + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} + \varepsilon_{Personal-sexual} \\
\eta_{Legal-illegal} &= \alpha_{Legal-illegal} + \beta_{14} x_{14} + \beta_{15} x_{15} + \beta_{16} x_{16} + \beta_{17} x_{17} + \beta_{18} x_{18} + \beta_{19} x_{19} + \varepsilon_{Legal-illegal}
\end{aligned} \tag{6}$$

Note: The sub-script indices numbers 1–19 correspond with the item numbers presented in Table 2.

Next, configural invariance was examined by specifying that the baseline model applies in all 44 countries [R Script: Configural Model].

$$\begin{aligned}
\eta^j_{Honesty-dishonesty} &= \alpha^j_{Honesty-dishonesty} + \beta^j_1 x^j_1 + \beta^j_2 x^j_2 + \beta^j_3 x^j_3 + \beta^j_4 x^j_4 + \beta^j_5 x^j_5 + \varepsilon^j_{Honesty-dishonesty} \\
\eta^j_{Personal-sexual} &= \alpha^j_{Personal-sexual} + \beta^j_6 x^j_6 + \beta^j_7 x^j_7 + \beta^j_8 x^j_8 + \beta^j_9 x^j_9 + \beta^j_{10} x^j_{10} + \beta^j_{11} x^j_{11} + \beta^j_{12} x^j_{12} + \beta^j_{13} x^j_{13} + \varepsilon^j_{Personal-sexual} \\
\eta^j_{Legal-illegal} &= \alpha^j_{Legal-illegal} + \beta^j_{14} x^j_{14} + \beta^j_{15} x^j_{15} + \beta^j_{16} x^j_{16} + \beta^j_{17} x^j_{17} + \beta^j_{18} x^j_{18} + \beta^j_{19} x^j_{19} + \varepsilon^j_{Legal-illegal}, \text{ for } j = 1 \text{ to } 44
\end{aligned} \tag{7}$$

Note:The sub-script indices numbers 1–19 correspond with the item numbers presented in Table 2.

Following, metric invariance was tested by keeping the specifications outlined in Eq. 7 and restricting all factor loadings to be as well equal across countries (Eq. 8) [R Script: Metric Model].

$$\beta^j_n = \beta^k_n, \text{ for all } n = 1, \ldots, 19 \text{ and all } j, k = 1, \ldots, 44 \tag{8}$$

Lastly, to verify scalar invariance, in addition to the previous model specifications (Eq. 7 and Eq. 8) the intercepts were restricted to be equal in all countries (Eq. 9) [R Script: Scalar Model].

$$\begin{aligned}\alpha^{j}_{Honesty-dishonesty} &= \alpha^{k}_{Honesty-dishonesty}\\ \alpha^{j}_{Personal-sexual} &= \alpha^{k}_{Personal-sexual}\\ \alpha^{j}_{Legal-illegal} &= \alpha^{k}_{Legal-illegal}\text{, for } \quad j, \quad k = 1, \ldots, 44\end{aligned} \tag{9}$$

If the fit of either the configural, metric, or scalar model was found to not be acceptable, or if a more restrictive model resulted in a significant decrease in model fit (as compared to a less restrictive model), the subsequent tests were not pursued. Although the $\chi 2$ test was reviewed, given the large sample size, we relied primarily on alternative fit indices—the comparative fit index (CFI; Bentler 1990), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR; Hu and Bentler 1999)—to evaluate the baseline model fit and configural invariance. For the CFI, acceptable fit was defined by values > 0.90; for the RMSEA, values < 0.10 and for the SRMR, values < 0.08 were considered acceptable. To determine whether metric or scalar invariance was confirmed, the fit of a more restrictive model was compared with that of the more liberal one. That is, the metric model was compared to the configural, and the scalar model was compared to the metric model. Here, in line with previous suggestions, changes in alternative fit indices were reviewed (Van de Schoot et al. 2012). We concluded metric noninvariance if $\Delta CFI \geq -0.01$ and $\Delta RMSEA \geq 0.015$ or $\Delta SRMR \geq 0.03$ were identified. Scalar noninvariance was determined when the comparison between the scalar and metric model yielded $\Delta CFI \geq -0.01$ and $\Delta RMSEA \geq 0.015$ or $\Delta SRMR \geq 0.01$ (Chen 2007; Cheung and Rensvold 2002; Rutkowski and Svetina 2014).

The aforementioned protocol is a stand-alone analysis, to be conducted in the step-by-step manner before any further analyses are carried out. MI tests should not be conducted by restricting the measurement models of latent variables in a structural equation model according to the requirements of either metric, scalar, or uniqueness invariance, such as by fixing factor loadings to be equal across groups. As described above, it is not absolute fit indices but changes in different alternative fit indices that are used to determine invariance above the configural level. This information cannot be attained if only one measurement model is fitted in the analysis.

Results

Descriptive Findings

Usually, this section would entail a presentation of the manifest mean scores and standard deviations of the three sub-scales of the MDBS for all countries. However, documenting this information would invite a comparison of manifest scores that is only justified if full measurement variance, that is, uniqueness invariance, has been documented. As will be shown below, such equivalence is not justified. Interested readers will find the manifest mean scores in the Supplementary Material (S2), with the note to not interpret any observed between-country variation.

Testing MI in the Full Sample (44 Countries; N = 59,482)

We first determined the fit of a three-factor baseline measurement model as described in Table 2 in the full sample. This model achieved acceptable fit after post-hoc model modification, specifically, after introducing additional covariances between the items 'Divorce'

and 'Sex before Marriage' as well as between the items 'Claiming government benefits to which you are not entitled' and 'Avoiding a fare on public transport' ($\chi 2$ (147) = 46,139.17, $p = 0.000$, CFI = 0.92, RMSEA = 0.073 (90%CI [0.072, 0.073]), SRMR = 0.07; Fig. 2).

As we had to rely on data-driven modification indices to attain acceptable model fit, we must highlight that the resulting baseline model is perhaps not applicable to other samples. Modification indices as reported when using lavaan or other similar software suggest the change in $\chi 2$ that can be attained when certain parameters are added to or omitted from a model. However, the use of those indices, although common, is criticized (MacCallum 1986). First, the suggestions are not informed by theory. Up and foremost, researchers must, therefore, only choose model modifications that are meaningful. The two correlations that we added reflect, in our opinion, sensible associations between items that speak to similar life circumstances (e.g., marriage) or settings. In addition, MacCallum et al. (1992) showed that modification indices are unstable in small and moderate-large samples. In other words, modifications that are identified in a sample are strongly informed by sample characteristics and not necessarily generalizable to the population. As we worked with a very large dataset, this concern is less pronounced (MacCallum et al. 1992). Working with smaller samples, researchers may opt for a cross-validation of the modified model in a second independent sample (see Browne and Cudeck 1989). Lastly, fewer modifications that result in larger $\chi 2$ changes are preferred as they are expected to be more stable than numerous modifications that attain each only small $\chi 2$-improvement (MacCallum et al. 1992). Hence, we selected the two modifications with the largest parameter change and did not add further modifications once acceptable model fit was attained.

Examining the baseline model across all countries – testing configural invariance – fit indices did not achieve the acceptable thresholds (Table 3). Following Rutkowski and Svetina's (2014) suggestion this decision was based on considering the CFI score (even if RMSEA and SRMR values were below the expected threshold). As configural invariance was not confirmed, it was not suitable to explore metric or scalar invariance (Cheung and Rensvold 2002). In other words, responding to Research Question 1, it is to be expected that there are countries in the dataset for which either the number of latent factors and/or the pattern of loadings of manifest variables on latent variables differs from the prescribed model.

Before moving on to the next analyses, we want to add a note on how to report results of MI tests. In Table 3 (as well as Tables 4, 5, 6, 7), we follow the recommendations set out by Putnick and Bornstein (2016). We believe that this format offers an accessible overview of all relevant information and encourage others to apply it as well.

Testing MI in Sub-Samples

Examining Research Question 2a and 2b, we next assessed MI in a selection of sub-samples of countries that scholars may intuitively assume to be similar or for while previous research identified commonalities. In all sub-sample analyses, we considered the baseline model (Fig. 2), including the covariances between the items 'Divorce' and'Sex before Marriage' as well as between the items'Claiming government benefits to which you are not entitled' and'Avoiding a fare on public transport'.

South America (Bolivia, Argentina, Brazil, Peru, Ecuador, Colombia, Chile; N = 7909) We first tested the baseline model in seven South American countries. The CFI was below

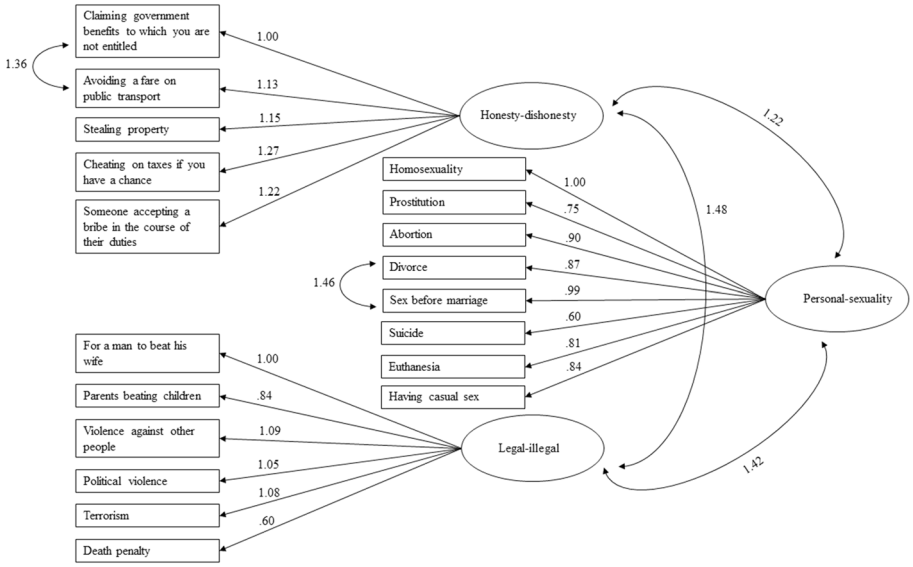

Fig. 2 Baseline model

the threshold; model fit was not considered acceptable ($\chi 2(147) = 6629.17$, $p = 0.000$; CFI = 0.88, RMSEA = 0.075, CI90% [0.073, 0.076]; SRMR = 0.07). As in the overall sample, in this sub-sample, the expected model of three latent variables, defined by the respective manifest variables outlines in Table 2, was not confirmed. No further analyses were conducted.

South-east Asia (Thailand, Myanmar, Singapore, Indonesia, Malaysia, Vietnam; N = 10,157) Next, we examined the baseline model in six South-east Asian countries. Here, the baseline model was supported ($\chi 2(147) = 7229.56$, $p = 0.000$; CFI = 0.93, RMSEA = 0.069, CI90% [0.068, 0.070]; SRMR = 0.06) as was configural invariance (Table 4). However, the CFI as well as ΔCFI values of the metric model were not satisfactory, such that metric noninvariance had to be concluded (Table 4).

East Asia (Hong Kong, Japan, Taiwan, Macao, South Korea, China; N = 9426) For six East Asian countries, the baseline model was not fully supported ($\chi 2(147) = 9485.99$, $p = 0.000$; CFI = 0.89, RMSEA = 0.082, CI90% [0.081, 0.084]; SRMR = 0.09) as the CFI value was below the acceptable threshold. No further analyses were completed.

Australasia (Australia, New Zealand; N = 2389) In the two Australasian countries, the baseline model fit was satisfactory ($\chi 2(147) = 1894.29$, $p = 0.000$; CFI = 0.91, RMSEA = 0.071, CI90% [0.068, 0.073]; SRMR = 0.06) and scalar MI was identified (Table 5).

North America (United States, Canada; N = 6461) Similarly, in the United States and Canada, the baseline model achieved acceptable fit ($\chi 2(147) = 5429.04$, $p = 0.000$; CFI = 0.92, RMSEA = 0.075, CI90% [0.073, 0.076]; SRMR = 0.086). In addition, scalar invariance was supported (Table 6).

Table 3 MI results full sample

Model	$\chi 2$ (df)	CFI	RMSEA (90% CI)	SRMR	Model comparison	$\Delta\chi 2$ (Δ df)	Δ CFI	Δ RMSEA	Δ SRMR
Configural	71,060.43 (6468)	0.88	0.086 (0.085, 0.087)	0.07					
Metric	–	–	–	–	Configural	–	–	–	–
Scalar	–	–	–	–	Metric	–	–	–	–

Table 4 MI results South-east Asia

Model	$\chi 2$ (df)	CFI	RMSEA (90% CI)	SRMR	Model comparison	$\Delta\chi 2$ (Δ df)	Δ CFI	Δ RMSEA	Δ SRMR
Configural	10,011.73 (882)	0.91	0.078 (0.077, 0.080)	0.06					
Metric	11,282.41 (962)	0.89	0.080 (0.078, 0.081)	0.08	Configural	1270.68 (80)	−0.02	0.002	0.02
Scalar	–	–	–	–	Metric	–	–	–	–

Table 5 MI results Australasia

Model	$\chi 2$ (df)	CFI	RMSEA (90% CI)	SRMR	Model comparison	$\Delta\chi 2$ (Δdf)	Δ CFI	Δ RMSEA	Δ SRMR
Configural	2150.15 (294)	0.91	0.073 (0.070, 0.076)	0.063					
Metric	2225.34 (310)	0.91	0.072 (0.069 0.075)	0.067	Configural	75.19 (16)	0.00	0.001	0.004
Scalar	2357.85 (326)	0.90	0.072 (0.070, 0.075)	0.068	Metric	132.51 (16)	0.01	0.00	0.001

Table 6 MI results North America

Model	$\chi 2$ (df)	CFI	RMSEA (90% CI)	SRMR	Model comparison	$\Delta\chi 2$ (Δdf)	Δ CFI	Δ RMSEA	Δ SRMR
Configural	5888.82 (294)	0.91	0.077 (0.075, 0.078)	0.085					
Metric	6000.12 (310)	0.91	0.075 (0.074 0.077)	0.087	Configural	111.3 (16)	0.00	− 0.002	0.002
Scalar	6560.93 (324)	0.90	0.077 (0.075, 0.079)	0.088	Metric	560.81 (14)	0.01	0.002	0.001

Table 7 MI results Anglophone countries

Model	$\chi 2$ (df)	CFI	RMSEA (90% CI)	SRMR	Model comparison	$\Delta\chi 2$ (Δdf)	Δ CFI	Δ RMSEA	Δ SRMR
Configural	8038.97 (588)	0.91	0.076 (0.074, 0.077)	0.09					
Metric	8404.51 (636)	0.91	0.074 (0.073 0.076)	0.08	Configural	365.54 (48)	0.00	0.002	0.004
Scalar	10,091.11 (684)	0.89	0.079 (0.077, 0.080)	0.09	Metric	1686.60 (48)	0.02	0.005	0.003

Anglophone Countries (United States, Canada, New Zealand, Australia; N = 8850). Lastly, for the sub-sample including all four Anglophone countries, the baseline model was acceptable ($\chi 2(147) = 6749.03$, $p = 0.000$; CFI = 0.92, RMSEA = 0.071, CI90% [0.070, 0.073]; SRMR = 0.08) as was the configural model (Table 7). The metric model fit was also satisfactory and the fit indices change scores confirmed metric invariance. However, considering ΔCFI and CFI scores, scalar invariance was not supported (Table 7).

Discussion

This article was motivated by three observations. First, cross-national criminological (survey) studies are flourishing as the field adopts an international perspective (Barberet 2007; Messner 2021). Second, accurate conclusions about associative relationships or between-group differences in cross-national datasets require metric or scalar invariance of the chosen (multi-item) scales (e.g., Byrne and Watkins 2003; Putnick and Bornstein 2016; Van de Schoot et al. 2012). Third, although measurement invariance has been examined in several cross-national studies, the implementation of the test is inconsistent and not always done such that relevant conclusions can be drawn (Schmitt and Kuljanin 2008; Vandenberg and Lance 2000). To promote the rigorous and wider adoption of MI tests, we aimed to provide an accessible 'step-by-step' protocol, accompanied by an analytical script, that documents how measurement invariance of multi-item scales can be assessed. Illustrating the analytical approach, we explored MI of the Morally Debatable Behavior Scale (Harding and Phillips 1986), fielded in wave seven of the WVS. Results showed that the three-factor structure of the scale, that was proposed in previous research, was not supported across the 44 countries in which the MDBS was administered (i.e., configural noninvariance). Consequently, without further adjustments (discussed in more detail below), estimates of cross-group comparisons of latent or manifest means of the MDBS as well as predictive associations, including those of multi-level models that draw on an aggregated dataset, are expected not to be accurate in the full sample.

Additionally, in sub-sets of countries that are in close proximity or that are thought to be culturally similar, metric and scalar invariance were not consistently achieved. For example, for six South-east Asian countries, we demonstrated configural invariance. However, the relative importance of each item for the respective sub-scales of the MDBS differed between countries (i.e., metric noninvariance). Conversely, in the sub-sample of four Anglophone countries, there was evidence for configural and metric invariance but intercepts of the measurement model were only equal across groups in the Australasian and North American countries respectively (i.e., scalar invariance). That is, latent scores of samples from Canada and the United States as well as Australia and New Zealand respectively can be confidently compared or aggregated.

First, and extending beyond the technical aspects of measurement invariance, the findings contribute to a literature that explores moral systems across cultures. The MDBS examines what behavior individuals consider to be justified, thus, it captures conceptions of morality. Principles of moral universalism, suggesting, for instance, universal moral values (Schwartz 1992), have been discussed for several decades (Haidt 2008). Although there is evidence for both a universalist and cultural-relativist perspective, Vauclair and colleagues (2014) emphasized that a combined approach might be most suitable. Specifically, based on the analysis of lay people's definition of morality they showed that for certain

themes (i.e., justice and welfare) a shared understanding of moral character can be found across cultures; the importance of right- and duty-based attributes, however, varied (Vauclair et al. 2014). The present study challenges those insights and the position that morality, as measured by the MDBS, is universal. In fact, we conclude that culture shapes the basic structure of what is considered moral as well as the extent to which specific activities define morally debatable behavior.

Our results also differed from outcomes of previous assessments of measurement invariance of other versions of the MDBS. Vauclair and Fischer's study (2011), which considered only the honesty-dishonest and personal–sexual sub-scales, did identify metric invariance in an earlier global WVS dataset. Moors and Wennekers (2003), focusing on a smaller sample of European countries, rejected any cross-country comparisons. Taken together, MI tests are essential in all data analysis plans of cross-national studies even when previous research suggests that countries might be culturally similar, when established scales are used, and also when a scale has demonstrated certain levels of invariance in other cross-national samples.

The fact that neither metric nor scalar invariance was confirmed in most of the samples that we investigated is not unusual (Davidov et al. 2014; Sokolov 2021; Welzel and Inglehart 2016). Indeed, it is likely that readers who apply our procedure to their own cross-national data will attain noninvariance. This might seem frustrating. However, numerous approaches can be applied to deal with a lack of adequate equivalence. First, detecting noninvariance of a scale is a valuable finding in its own right as it can point to cultural over-generalization (Fischer and Karl 2019). For instance, our analyses demonstrated that morally debatable behaviors were conceptualized differently between the South American and East-asian countries that were studied. In the South-east Asian countries, the overall configuration of morally debatable behaviors was comparable but, in some settings, certain indicator were more relevant for defining particular sub-scales. These insights are a starting point for further research, notably, qualitative interviews, that could aid to explore the differences (see Lugtig et al. (2011) for an example that used this approach to identify changes in students' certainty about their study motivation over time).

Additionally, once noninvariance is detected, exploratory analyses can be conducted to identify the specific items that are not invariant (Cheung and Rensvold 2002; Sokolov 2021). Cheung and Rensvold (2002), based on Byrne and colleagues (1989), suggested that for a multi-factor scale such as the MDBS (i.e., honesty-dishonesty, personal-sexual, legal-illegal sub-scales), loadings associated with one sub-scale must be fixed to be equal across groups, while all other loadings are permitted to vary. This process is repeated for all sub-scales. Fit of all models is compared to the fully restricted model to identify noninvariant latent constructs. Next, separate models are estimated for each item in each noninvariant construct, restricting one item at a time to be equal across groups until the noninvariant items are detected. Vijver and Leung (1997) proposed to detect the source of noninvariance based on univariate analyses of variance that examines whether specific item scores (the dependent variable) are more (or less) likely in any one sub-group. Finally, Fischer and Karl (2019) introduced an accessible way to identify metric invariant items using R, which can be combined with the analytical code that accompanies this paper.

It is important to note that if at least two factor loadings and intercepts are equal across groups, analyses may, in fact, be conducted while accepting partial (metric or scalar) invariance (Byrne et al. 1989). Pokropek and colleagues (2019) showed in a simulation study with a five-item scale that even if up to 80% of items are noninvariant, mean scores and factor loadings are still estimated accurately. Interestingly, the authors highlighted that scales with only three or four items are more affected, that is, accurate estimates are harder

to compute under conditions of noninvariance. Having said this, and as no guidelines are available regarding the acceptable level of non-equivalence, the decision to allow for partial invariance is best informed by an understanding of its impact (see Meuleman 2012). Schmitt and colleagues (2011) illustrated an approach for doing so. Specifically, they tested if and to what extent estimates of regression coefficients and latent mean scores differ between a) a model where noninvariance is ignored and factor loadings as well as intercepts are expected to be equal across study units and b) a model in which certain parameters are allowed to vary between groups.

Alternatively, to cope with noninvariance in a dataset as diverse as the global sample that we used, mixture multigroup factor analysis can be conducted (De Roover 2021). Mixture multigroup factor analysis is a data-driven method that aims to identify sub-samples, or classes, for which certain model parameters are equivalent. As classes are defined, only a specified set of parameters is fixed to be equivalent within the sub-sets. For instance, if factor loadings are fixed to be invariant within classes, intercepts are still permitted to be variant for the same study units. It is recommended to apply the mixture multigroup factor analysis in steps, exploring classes where loadings are equivalent and then clusters were intercepts are invariant; else, one makes the assumption that the same logic underlies each level of equivalence (Leitgöb et la., 2022). De Roover (2021) offer guidance on how to conduct the analysis as well as the relevant R package and script that can be combined with our script.

Lastly, it must be noted that there are less strict methods available to test MI. Specifically, Bayesian structural equation models can be used to assess approximate invariance (Muthén and Asparouhov 2012). Here, restrictions on factor loadings or intercepts are not set to exactly zero but 'approximately zero', that is, the scores are expected to vary around zero with pre-determined (larger or smaller) variance (i.e., the prior), reflecting a normal distribution (van de Schoot et al. 2012). A balance is sought between accepting a degree of non-substantive parameter-differences while still striving for optimal model fit. Schoot et al. (2012) demonstrated that restrictions of approximate measurement invariance are more appropriate than those imposed by traditional approaches, such as the multi-group confirmatory factor analysis that was introduced in this paper, when there are small differences between groups on many intercepts. Leitgöb and colleagues (2022) offer a valuable discussion on the selection of suitable priors.

Limitations

Although we aimed to provide substantial practical guidance for how to conduct MI tests in cross-national criminological research, this primer is not without its own limitation. First, we tested MI with multi-group confirmatory factor analysis. This procedure is widely applied, and we believe that it is most accessible for those who have never conducted MI analyses. However, other analytical approaches are available – namely, exploratory factor analysis (Meredith 1964), multidimensional scaling (Braun and Scott 1998), item response theory (Raju et al. 2002), and latent class analysis (see Millsap and Meredith (2007) for more detail) – and we cannot rule out that those would arrive at different conclusions. Future research should, in fact, explore in more detail whether and how the choice of analytical perspective affects the outcomes of MI tests.

It is further important to note that any assessment of MI, such as our analysis, does not provide definite insights about the reliability or validity of a scale. In principle, lower reliability of a scale in a study-unit can be an indicator of measurement noninvariance; in turn,

if configural MI is not achieved, it is to be expected that an instrument does not hold equal levels of reliability in all groups (Chen 2008). However, MI assessment cannot indicate if the respective reliabilities are acceptable. Likewise, whether an instrument captures what it intends to, validity, must be determined independently. Indeed, confirming the factor structure of a measure, and doing so in different sub-samples (i.e., configural invariance), pertains to certain phases of validity tests. Establishing convergent or discriminatory validity does demand its own analyses.

Moreover, any conclusions that we drew about the level of invariance in any assessed sub-sample should not be extended to countries in the same region that were not included in the analysis. For instance, considering the failure to replicate the scale factor structure in the South-american sub-sample, it is unclear whether MI tests based on data from Uruguay and Paraguay would yield the same results. Importantly, although we presented suggestions for how to overcome measurement invariance, we did not apply them in our analysis to identify a fully invariant MDBS in the sample of 44 countries. We encourage others to consider this step, making a valuable contribution to cross-national criminology. Furthermore, future studies.

Lastly, as noted, we removed all cases from the analysis that included at least one missing value, on any one item of the scale. As a result, we excluded whole countries (Table 1) from the analysis, which reduces the overall scope of the conclusions that can be drawn. Moreover, working only with complete cases means that individuals who might have had a different understanding of a particular item are perhaps systematically excluded. Answering 'refuse', in particular, could imply that those participants evaluated the respective behaviors as more severe or unacceptable than other participants who did provide an answer. As we could not confidently point to any reasonable variable that could have explained the missing data patterns (i.e., missing at random), we considered list wise exclusion the more prudent approach. Missing data patterns should always be inspected in cross-national analyses. If missing completely at random or missing at random patterns are identified, the analytical procedure that we described can be conducted with case-wise (or 'full information') maximum likelihood estimation in R (Rosseel, 2012c).

Conclusions

In recent years, large-*n* cross-national surveys have become increasingly relevant in criminological research. Measurement noninvariance of multi-item instruments that are employed in these studies could, however, challenge the accuracy and robustness of results. We highlight that MI cannot simply be assumed but must be empirically assessed; we also present a step-by-step analytical procedure documenting how this can be practically done. To encourage the adoption of MI tests in criminological research, and perhaps even greater interest in conducting impactful cross-national studies in general, we conclude this paper by summarizing eight recommendations:

1. First, measurement invariance of multi-item instruments that examine an underlying latent construct must be examined whenever data from two or more groups is included in an analysis – either to compare mean scores or regression coefficients, or to aggregate sub-samples. The present research has focused on cross-national studies. However, noninvariance might also be a concern when considering sub-samples defined by different gender or ethnicity. Moreover, and although not discussed in this paper, MI tests are

required in longitudinal designs to verify whether the conceptualization of a construct changes over time (see Adolf et al. 2014).
2. Data that is used in MI tests should be collected in the same manner in all sub-groups.
3. The instruments for which equivalence is examined must be designed in the same way in all groups. Specifically, the answer options (i.e., steps on a Likert-type scale, adding numeric values to text labels) must be identical (Leitgöb et al. 2022).
4. A sample size of $n > 200$ is recommended per sub-group (Fischer and Karl 2019).
5. Multi-group confirmatory factor analysis that we introduced in this paper, and for which we provide an annotated R Script and output file, is a common procedure to test measurement invariance. Configural, metric, and scalar invariance are tested in a step-by-step process. Although we believe that this protocol is most accessible, especially for those new to MI tests, we invite readers to explore in particular Bayesian/approximate and partial invariance as they have been found to be equally informative in certain scenarios (Pokropek et al. 2019; Schoot et al. 2012).
6. When determining the baseline measurement model, which is then fitted in all sub-groups, modification indices could be used to identify which parameters might be changed to improve model fit. We recommend being very cautious when working with modification indices. Implementation of the suggestions should be based on theoretical reasoning. Ideally, fewer alterations that result in the largest model fit improvement are introduced.
7. Putnick and Bornstein (2016) provide a template to report the results of MI tests. Although other formats may be applied as well, we find the proposed structure of tables (see Table 3–7 in this paper) especially helpful to give readers a concise overview of findings.
8. Lastly, noninvariance is not an unusual outcome of MI tests. Therefore, researchers should reflect in advance on how they may choose to deal with this situation. Different approaches are alluded to in this paper. However, it is important to consider a result of nonequivalence as valuable in its own right, and possibly explore follow up research that examines the source of the identified differences.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10940-023-09578-9.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Adolf J, Schuurman NK, Borkenau P, Borsboom D, Dolan CV (2014) Measurement invariance within and between individuals: A distinct problem in testing the equivalence of intra-and inter-individual model structures. Front Psychol 5:883

Aebi MF, Linde A (2015) The epistemological obstacles in comparative criminology: a special issue introduction. Eur J Criminol 12(4):381–385. https://doi.org/10.1177/147737081559
Barberet R (2007) The internationalization of criminology? A content analysis of presentations at American Society of Criminology Conferences. J Crim Just Educ 18(3):406–427. https://doi.org/10.1080/10511250701705362
Bennett RR (1980) Constructing cross-cultural theories in criminology: application of the generative approach. Criminology 18(2):252–268. https://doi.org/10.1111/j.1745-9125.1980.tb01364.x
Bennett RR (2004) Comparative criminology and criminal justice research: the state of our knowledge. Justice Q 21(1):1–21. https://doi.org/10.1080/07418820400095721
Bennett RR (2009) Comparative criminological and criminal justice research and the data that drive them. Int J Comp Appl Crim Just 33(2):171–192. https://doi.org/10.1080/01924036.2009.9678804
Bentler PM (1990) Comparative fit indexes in structural models. Psychol Bull 107(2):238–246. https://doi.org/10.1037/0033-2909.107.2.238
Bieda A, Hirschfeld G, Schönfeld P, Brailovskaia J, Zhang XC, Margraf J (2017) Universal happiness? Cross-cultural measurement invariance of scales assessing positive mental health. Psychol Assess 29(4):408–421
Braun M, Scott J (1998) Multidimensional scaling and equivalence: is having a job the same as working?. 3: 129–144
Browne MW, Cudeck R (1989) Single sample cross-validation indices for covariance structures. Multivar Behav Res 24(4):445–455. https://doi.org/10.1207/s15327906mbr2404_4
Byrne BM, Shavelson RJ, Muthén B (1989) Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. Psychol Bull 105(3):456–466
Byrne BM, Watkins D (2003) The issue of measurement invariance revisited. J Cross Cult Psychol 34(2):155–175. https://doi.org/10.1177/002202210225
Chen FF (2007) Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct Equ Modeling: Multidiscip J 14(3):464–504
Chen FF (2008) What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. J Pers Soc Psychol 95(5):1005–1018. https://doi.org/10.1037/a0013193
Cheung GW, Rensvold RB (2002) Evaluating goodness-of-fit indexes for testing measurement invariance. Struct Equ Model 9(2):233–255. https://doi.org/10.1207/S15328007SEM0902_5
Chon DS (2021) Muslims, religiosity, and attitudes toward wife beating: analysis of the world values survey. Int Criminol 1(2):150–164. https://doi.org/10.1007/s43576-021-00016-z
Davidov E, Meuleman B, Cieciuch J, Schmidt P, Billiet J (2014) Measurement equivalence in cross-national research. Ann Rev Sociol 40:55–75. https://doi.org/10.1146/annurev-soc-071913-043137
De Roover K (2021) Finding clusters of groups with measurement invariance: unraveling intercept noninvariance with mixture multigroup factor analysis. Struct Equ Modeling 28(5):663–683. https://doi.org/10.1080/10705511.2020.1866577
Diamantopoulos A, Sarstedt M, Fuchs C, Wilczynski P, Kaiser S (2012) Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective. J Acad Mark Sci 40:434–449. https://doi.org/10.1007/s11747-011-0300-3
Dong Y, Dumas D (2020) Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. Personal Individ Differ 160:109956. https://doi.org/10.1016/j.paid.2020.109956
Fischer R, Karl JA (2019) A primer to (cross-cultural) multi-group invariance testing possibilities in R. Front Psychol. https://doi.org/10.3389/fpsyg.2019.01507
Franke GR, Nadler SS (2008) Culture, economic development, and national ethical attitudes. J Bus Res 61(3):254–264. https://doi.org/10.1016/j.jbusres.2007.06.005
Gerstner D, Wickes R, Oberwittler D (2019) Collective efficacy in Australian and German neighborhoods: Testing cross-cultural measurement equivalence and structural correlates in a multi-level SEM framework. Soc Indic Res 144(3):1151–1177. https://doi.org/10.1007/s11205-019-02081-4
Guenole N, Brown A (2014) The consequences of ignoring measurement invariance for path coefficients in structural equation models. Front Psychol. https://doi.org/10.3389/fpsyg.2014.00980
Haerpfer C, Inglehart R, Moreno A, Welzel C, Kizilova K, Diez-Medrano J, Lagos M, Norris P, Ponarin E, Puranen B, et al. (eds.) (2020) World values survey: round seven – country-pooled datafile. JD Systems Institute & WVSA Secretariat, Madrid. https://doi.org/10.14281/18241.1
Haerpfer C, Inglehart R, Moreno A, Welzel C, Kizilova K, Diez-Medrano J, et al. (2022) World values survey wave 7 (2017–2022) Cross-National Data-Set. World Values Survey Association
Haidt J (2008) Morality. Perspect Psychol Sci 3:65–72

Halpern D (2001) Moral values, social trust and inequality: can values explain crime?. Brit J Criminol 41(2):236–251
Harding S, Phillips D (1986) Contrasting values in western Europe. Unity, diversify, and change. Macmillan, London
Herrero J, Torres A, Rodríguez FJ, Juarros-Basterretxea J (2017) Intimate partner violence against women in the European Union: the influence of male partners' traditional gender roles and general violence. Psychol Violence 7(3):385–394. https://doi.org/10.1037/vio0000099
Hirtenlehner H, Farrall S, Bacher J (2013) Culture, institutions, and morally dubious behaviors: testing some core propositions of the institutional-anomie theory. Deviant Behav 34(4):291–320. https://doi.org/10.1080/01639625.2012.726165
Hofstede G (2022) Hofstede insights. https://www.hofstede-insights.com/product/compare-countries/
Hu LT, Bentler PM (1999) Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 6(1):1–55. https://doi.org/10.1080/10705519909540118
James HS Jr (2015) Generalized morality, institutions and economic growth, and the intermediating role of generalized trust. Kyklos 68(2):165–196. https://doi.org/10.1111/kykl.12079
Jöreskog KG (1971) Simultaneous factor analysis in several populations. Psychometrika 36(4):409–426
Julkif NB (2022) Self and political efficacy and the justifiability of political violence and the role of state terror: a cross-national analysis. Soc Sci Q 103(1):108–119. https://doi.org/10.1111/ssqu.13120
Kafafian M, Botchkovar EV, Marshall IH (2022) Moral rules, self-control, and school context: additional evidence on situational action theory from 28 Countries. J Quant Criminol 38(4):861–889. https://doi.org/10.1007/s10940-021-09503-y
Karstedt S (2001) Comparing cultures, comparing crime: challenges, prospects and problems for a global criminology. Crime Law Soc Chang 36(3):285–308. https://doi.org/10.1023/A:1012223323445
Katz RC, Santman J, Lonero P (1994) Findings on the revised morally debatable behaviors scale. J Psychol 128(1):15–21. https://doi.org/10.1080/00223980.1994.9712707
Kovandzic T, Kleck G (2022) The impact of firearm levels on homicide rates: the effects of controlling for cultural differences in cross-national research. Am J Crim Justice 47(1):41–55. https://doi.org/10.1007/s12103-020-09604-7
LaFree G (2021) Progress and obstacles in the internationalization of criminology. Int Criminol 1(1):58–69. https://doi.org/10.1007/s43576-021-00005-2
Leitgöb H, Seddig D, Asparouhov T, Behr D, Davidov E, De Roover K, van de Schoot R (2022) Measurement invariance in the social sciences: historical development, methodological challenges, state of the art, and future perspectives. Soc Sci Res. https://doi.org/10.1016/j.ssresearch.2022.102805
Letki N (2006) Investigating the roots of civic morality: trust, social capital, and institutional performance. Polit Behav 28(4):305–325. https://doi.org/10.1007/s11109-006-9013-6
Lugtig PJ, Boeije HR, Lensvelt-Mulders GJLM (2011) Change? What change? An exploration of the use of Mixed-methods research to understand longitudinal measurement invariance. Methodology 8(4):1–9
MacCallum R (1986) Specification searches in covariance structure modeling. Psychol Bull 100(1):107–120. https://doi.org/10.1037/0033-2909.100.1.107
MacCallum RC, Roznowski M, Necowitz LB (1992) Model modifications in covariance structure analysis: the problem of capitalization on chance. Psychol Bull 111(3):490504. https://doi.org/10.1037/0033-2909.111.3.490
Marozzi M (2021) Perceived justifiability towards morally debatable behaviors across Europe. Soc Indic Res 153(2):759–778. https://doi.org/10.1007/s11205-020-02490-w
Marshall IH, Birkbeck C, Enzmann D, Kivivuori J, Markina A, Steketee M (2022) International self-report delinquency (ISRD4) study protocol: background, methodology and mandatory items for the 2021/2022 survey. Northeastern University, Boston
Martínez PR, Sánchez AJS, Galindo CJA (2022) Justification of terrorism according to World Values Survey (2017–2020). Res Glob. https://doi.org/10.1016/j.resglo.2022.100085
Meredith W (1993) Measurement invariance, factor analysis and factorial invariance. Psychometrika 58(4):525–543. https://doi.org/10.1007/BF02294825
Meredith W (1964) Notes on factorial invariance. Psychometrika 29(2):177–185. https://doi.org/10.1007/BF02289699
Messner SF (2015) When west meets east: generalizing theory and expanding the conceptual toolkit of criminology. Asian J Criminol 10(2):117–129. https://doi.org/10.1007/s11417-014-9197-3
Messner SF (2021) The glass is at least half full: Reflections on the internationalization of criminology. Int Criminol 1(1):13–19. https://doi.org/10.1007/s43576-020-00001-y

Meuleman B (2012) When are item intercept differences substantively relevant in measurement invariance testing?. In: Methods, theories, and empirical applications in the social sciences, VS Verlag für Sozialwissenschaften, pp 97–104

Millsap RE, Meredith W (2007) Factorial invariance: Historical perspectives and new problems. Factor analysis at 100. Routledge, London, pp 145–166

Moors G, Wennekers C (2003) Comparing moral values in Western European countries between 1981 and 1999. A multiple group latent-class factor approach. Int J Comp Sociol 44(2):155–172. https://doi.org/10.1177/002071520304400203

Muthén B, Asparouhov T (2012) Bayesian structural equation modeling: a more flexible representation of substantive theory. Psychol Methods 17(3):313–335. https://doi.org/10.1037/a0026802

Nivette AE (2021) Exploring the availability and potential of international data for criminological study. Int Criminol 1(1):70–77. https://doi.org/10.1007/s43576-021-00009-y

Nivette A, Trajtenberg N, Eisner M, Ribeaud D, Peres MFT (2020) Assessing the measurement invariance and antecedents of legal cynicism in São Paulo, Zurich, and Montevideo. J Adolesc 83:83–94. https://doi.org/10.1016/j.adolescence.2020.06.007

Pauwels L, Pleysier S (2005) Assessing cross-cultural validity of fear of crime measures through comparisons between linguistic communities in Belgium. Eur J Criminol 2(2):139–159.

Pechorro P, DeLisi M, Pacheco C, Abrunhosa Gonçalves R, Maroco J, Quintas J (2022) Examination of Grasmick et al.'s low self-control scale and of a short version with cross-gender measurement invariance. Crime Delinquency. https://doi.org/10.1177/00111287211073674

Pleysier S, Vervaeke G, Goethals L (2004) Cross-cultural invariance and gender bias when measuring 'fear of crime.' Int Rev Victimol 10(3):245–260. https://doi.org/10.1177/026975800401000303

Pokropek A, Davidov E, Schmidt P (2019) A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. Struct Equ Model 26(5):724–744. https://doi.org/10.1080/10705511.2018.1561293

Putnick DL, Bornstein MH (2016) Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. Dev Rev 41:71–90. https://doi.org/10.1016/j.dr.2016.06.004

Raju NS, Laffitte LJ, Byrne BM (2002) Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. J Appl Psychol 87(3):517–529. https://doi.org/10.1037/0021-9010.87.3.517

Rodríguez JA, Pérez-Santiago N, Birkbeck C (2015) Surveys as cultural artefacts: applying the international self-report delinquency study to Latin American adolescents. Eur J Criminol 12(4):420–436. https://doi.org/10.1177/1477370815581701

Rogers ML, Pridemore WA (2022) Not Just another test of institutional anomie theory: assessing relative institutional imbalances. Justice Quart. https://doi.org/10.1080/07418825.2022.2102535

Rosseel Y (2012) lavaan: an R package for structural equation modeling. J Stat Softw 48(2):1–36. https://doi.org/10.18637/jss.v048.i02

Rosseel Y (2023). Lavaan tutorial, estimators. Available at: https://lavaan.ugent.be/tutorial/est.html

Rutkowski L, Svetina D (2014) Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. Educ Psychol Measur 74(1):31–57. https://doi.org/10.1177/0013164413498257

Sampson RJ, Raudenbush SW, Earls F (1997) Neighborhoods and violent crime: a multilevel study of collective efficacy. Science 277(5328):918–924. https://doi.org/10.1126/science.277.5328.9

Schmitt N, Golubovich J, Leong FT (2011) Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: an illustrative example using Big Five and RIASEC measures. Assessment 18(4):412–427. https://doi.org/10.1177/1073191110373223

Schmitt N, Kuljanin G (2008) Measurement invariance: review of practice and implications. Hum Resour Manag Rev 18(4):210–222. https://doi.org/10.1016/j.hrmr.2008.03.003

Schwartz SH (1992) Universals in the content and structure of values: theoretical advances and empirical tests in 20 countries. Adv Exp Soc Psychol 25:1–65. https://doi.org/10.1016/S0065-2601(08)60281-6

Sokolov B (2021) Measurement invariance of liberal and authoritarian notions of democracy: evidence from the world values survey and additional methodological considerations. Front Polit Sci. https://doi.org/10.3389/fpos.2021.642283

Spencer MS, Fitch D, Grogan-Kaylor A, Mcbeath B (2005) The equivalence of the behavior problem index across US ethnic groups. J Cross Cult Psychol 36(5):573–589. https://doi.org/10.1177/0022022105278543

Steinmetz H (2013) Analyzing observed composite differences across groups: is partial measurement invariance enough? Methodol Eur J Res Methods Behav Soc Sci 9(1):1. https://doi.org/10.1027/1614-2241/a000049

Storm I (2016) Morality in context: a multilevel analysis of the relationship between religion and values in Europe. Polit Religion 9(1):111–138. https://doi.org/10.1017/S1755048315000899
Tausch A (2019) Multivariate analyses of the global acceptability rates of male intimate partner violence (IPV) against women based on W orld V alues S urvey data. Int J Health Plann Manage 34(4):1155–1194. https://doi.org/10.1002/hpm.2781
The DHS Program (2023) The demographic and health survey. https://dhsprogram.com/
Thulin EJ, Heinze JE, Kusunoki Y, Hsieh HF, Zimmerman MA (2021) Perceived neighborhood characteristics and experiences of intimate partner violence: a multilevel analysis. J Interpersonal Violence 36:23–24. https://doi.org/10.1177/0886260520906183
Tonry M (2015) Is cross-national and comparative research on the criminal justice system useful? Eur J Criminol 12(4):505–516. https://doi.org/10.1177/1477370815581699
Triandis HC (1978) Some universals of social behavior. Pers Soc Psychol Bull 4(1):1–16. https://doi.org/10.1177/014616727800400101
Van de Schoot R, Lugtig P, Hox J (2012) A checklist for testing measurement invariance. Eur J Dev Psychol 9(4):486–492. https://doi.org/10.1080/17405629.2012.686740
van de Vijver FJR (1998) Towards a theory of bias and equivalence. In ZUMA (Centrum fur Umfragen Methoden und Analysen)-Nachrichten Spezial Band 3: Cross-Cultural Survey Equivalence, pp 41–65. http://www.gesis.org/Publikationen/Zeitschriften/ZUMA_Nachrichten_spezial/zn-sp-3-inhalt.htm
Vandenberg RJ, Lance CE (2000) A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organ Res Methods 3(1):4–70. https://doi.org/10.1177/109442810031002
Van Dijk JJM, Mayhew P, Killias M (1990) Experiences of crime across the world: Key findings from the 1989 International Crime Survey. Kluwer Law and Taxation, Deventer
Vauclair CM, Fisher R (2011) Do cultural values predict individuals' moral attitudes? A cross-cultural multilevel approach. Eur J Soc Psychol 41(5):645–657. https://doi.org/10.1002/ejsp.794
Vauclair CM, Wilson M, Fischer R (2014) Cultural conceptions of morality: examining laypeople's associations of moral character. J Moral Educ 43(1):54–74
Van de Vijver F, Leung K (1997) Methods and data analysis of comparative research. Handbook of cross-cultural psychology. Allyn & Bacon, London, pp 257–300
Welkenhuysen-Gybels J, van de Vijver FJR, Cambré B (2007) A comparison of method for the evaluation of construct equivalence in a multigroup setting. Meas Mean Data Soc Res, 357–371
Welzel C, Inglehart RF (2016) Misconceptions of measurement equivalence: time for a paradigm shift. Comp Pol Stud 49(8):1068–1094. https://doi.org/10.1177/0010414016628275
Wicherts JM, Dolan CV, Hessen DJ (2005) Stereotype threat and group differences in test performance: a question of measurement invariance. J Pers Soc Psychol 89(5):696–716. https://doi.org/10.1037/0022-3514.89.5.696
Widaman KF, Reise SP (1997) Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In: Bryant KJ, Windle M, West SG (eds) The science of prevention: methodological advances from alcohol and substance abuse research. American Psychological Association, pp 281–324
Zito RC (2019) Institutional anomie and justification of morally dubious behavior and violence cross-nationally: a multilevel examination. Aust N Z J Criminol 52(2):250–271. https://doi.org/10.1177/0004865818785653

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.