

# Why Measurement Invariance is Important in Comparative Research. A Response to Welzel et al. (2021)

Sociological Methods & Research

2023, Vol. 52(3) 1401–1419

© The Author(s) 2022



Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/00491241221091755

[journals.sagepub.com/home/smr](https://journals.sagepub.com/home/smr)



**Bart Meuleman** <sup>1</sup>, **Tomasz Żóltak** <sup>2</sup>,  
**Artur Pokropek** <sup>3</sup>, **Eldad Davidov** <sup>4,5</sup>,  
**Bengt Muthén**<sup>6</sup>, **Daniel L. Oberski**<sup>7</sup>,  
**Jaak Billiet**<sup>8</sup>, and **Peter Schmidt**<sup>9</sup>

## Abstract

Welzel et al. (2021) claim that non-invariance of instruments is inconclusive and inconsequential in the field for cross-cultural value measurement. In this response, we contend that several key arguments on which Welzel et al.

<sup>1</sup>Centre for Sociological Research (CeSO), Institute for Social and Political Opinion Research, University of Leuven, Leuven, Vlaams-Brabant, Belgium

<sup>2</sup>Institute of Philosophy and Sociology, Polish Academy of Sciences, Warszawa, Mazowieckie, Poland

<sup>3</sup>Educational Research Institute (IBE), Warszawa, Mazowsze, Poland

<sup>4</sup>Institute of Sociology and Social Psychology, University of Cologne, Köln, Nordrhein-Westfalen, Germany

<sup>5</sup>Department of Sociology and URPP Social Networks, University of Zurich, Zurich, Switzerland

<sup>6</sup>Department of Education, UCLA, Los Angeles, California, USA

<sup>7</sup>Department of Methodology and Statistics, Utrecht University, Utrecht, Netherlands

<sup>8</sup>Centre for Sociological Research (CeSO), Institute for Social and Political Opinion Research, University of Leuven, Leuven, Vlaams-Brabant, Belgium

<sup>9</sup>Department of Political Sciences, University of Giessen, Giessen, Hessen, Germany

## Corresponding Author:

Bart Meuleman, Centre for Sociological Research (CeSO), Institute for Social and Political Opinion Research, University of Leuven, Leuven, Vlaams-Brabant, Belgium.

Email: [bart.meuleman@kuleuven.be](mailto:bart.meuleman@kuleuven.be)

(2021) base their critique of invariance testing are conceptually and statistically incorrect. First, Welzel et al. (2021) claim that value measurement follows a formative rather than reflective logic. Yet they do not provide sufficient theoretical arguments for this conceptualization, nor do they discuss the disadvantages of this approach for validation of instruments. Second, their claim that strong inter-item correlations cannot be retrieved when means are close to the endpoint of scales ignores the existence of factor-analytic approaches for ordered-categorical indicators. Third, Welzel et al. (2021) propose that rather than of relying on invariance tests, comparability can be assessed by studying the connection with theoretically related constructs. However, their proposal ignores that external validation through nomological linkages hinges on the assumption of comparability. By means of two examples, we illustrate that violating the assumptions of measurement invariance can distort conclusions substantially. Following the advice of Welzel et al. (2021) implies discarding a tool that has proven to be very useful for comparativists.

### **Keywords**

Measurement invariance, cross-cultural research, reflective vs. formative measurement, ordered-categorical data analysis, nomological linkages

### **Introduction**

In their recent paper, Welzel, Brunkert, Kruse, and Inglehart (2021) take issue with the practice of testing measurement invariance of multi-item instruments in cross-cultural research. Welzel et al. (2021) claim that violations of the invariance assumption are statistical artefacts that are inconsequential for making meaningful comparisons. The authors' critique follows a critical evaluation of their own cultural value scales (Alemán and Woods 2016; see also Sokolov 2018), and the argumentation echoes sentiments the authors published previously in other journals (Welzel and Inglehart 2016; Welzel, Brunkert, Inglehart and Kruse 2019).

We fully agree that scientists should be wary of dogmatism and prepared to discuss openly whether our practices – including invariance testing – are useful. In this sense, we welcome that Welzel et al. (2021) critically examine whether invariance testing achieves its stated aims. However, the authors misrepresent the current practice of invariance testing, leading to a

straw-man argument that discards the actual tool proven to be useful for comparativists.

The idea behind “measurement invariance” is simple and – presumably – uncontroversial: when we compare any measurement across groups, that comparison should reflect true differences rather than measurement differences. This idea – that comparability mandates comparable measures – is ancient, and can be found throughout history – including in the Magna Carta (1215). In modern surveys, we often use statistical models of the response process to represent the measurement, translating the old adage into a hypothesis that the process, or certain relevant model parameters that describe it, does not differ across groups (Mellenbergh 1989). It is exactly this hypothesis that measurement invariance testing aims to confront with data. For example, suppose we would like to compare two countries on their citizens’ levels of “conscientiousness”, measured by a survey. Then respondents who are from different countries but who have the same true level of “conscientiousness” should score similarly on the indicators measuring that trait (Davidov et al. 2014).

The hypothesis of comparable measurement has testable consequences for the observed interrelations among items of a multi-indicator instrument. The measurement invariance hypothesis states that these differences should, within sampling variability, follow a specific pattern dictated by the model (see also Meredith and Millsap 1992). Often this model is multiple-group confirmatory factor analysis (MGCFA; Jöreskog 1971; Millsap 2011), but psychometric work on differential item functioning (DIF) is also popular (for an overview see Holland and Wainer 2012). New models and approaches that test the conditions for making meaningful comparisons are continually being developed (see Davidov, Muthén, and Schmidt 2018, for a special issue in this journal on the topic). These innovations reflect a constant search for techniques that allow us to distinguish ignorable cross-group non-invariance from veritable sources of bias. The notions of partial invariance (Byrne, Shavelson, and Muthén 1989; Steenkamp and Baumgartner 1998), approximate measurement invariance (AMI), and partial approximate measurement invariance (PAMI; Asparouhov and Muthén 2014) are examples of the fruits of this search.

The picture Welzel et al. (2021) paint of measurement invariance testing requires correction. This response rebuts the main points of criticism Welzel et al. (2021) raise. We start by discussing the three main lines of argumentation of Welzel et al. (2021). Subsequently, we illustrate by means of realistic examples of research – including work of Welzel and

colleagues – how ignoring the assumption of invariance can lead to wrong conclusions.

### Three Main Arguments of Welzel et al. (2021)

Welzel et al. (2021) build their argument against the measurement invariance paradigm on three claims.

First, they argue that the indices of value dimensions often used in cross-cultural research follow the logic of formative measurement rather than reflective measurement, which puts them “out of reach of MGCFA’s judgmental authority” (Welzel et al. 2021:4). We will refer to this claim as “the argument of formative constructs.”

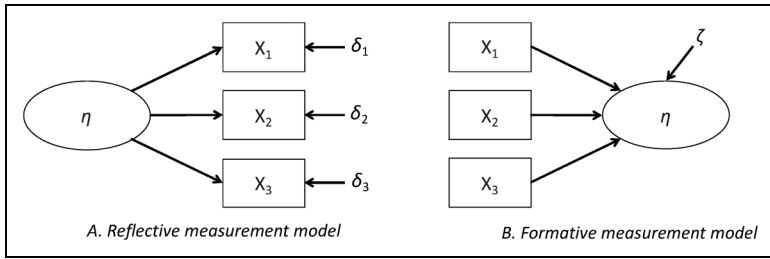
Second, they argue that statistical methods of testing for measurement invariance have no use in empirical cross-country comparative analysis because of the closed-ended nature of the measures of the scales. Furthermore, the authors claim “that the functioning of a *multi*-item index across groups depends in no way on the functioning of its single items within groups” (Welzel et al. 2021:4). Following the authors, we refer to this claim as the “argument of restricted realm of applicability.”

Third, the authors argue that nomological linkages at the aggregate level are sufficient as guiding principle for establishing measurement equivalence. In short, they claim that looking at relationships between country-level aggregates automatically evaluates comparability. We will refer to this argument as the “argument of nomological network linkages.”

We consider all three arguments as incorrect, straw-man, or exaggerated. Below, we elaborate why we – respectfully, but forcefully – disagree.

### *The Argument of Formative Constructs*

The first argument of Welzel et al. (2021:9) is that their scales measuring secular and emancipative values are based on a *combinatory* instead of a *dimensional* logic. This distinction is known in the methodological literature as the difference between “formative” and “reflective” constructs. Formative and reflective measurement indeed refer to two fundamentally different models of mapping theoretical constructs onto empirical observations. These two *auxiliary theories* of measurement start from different philosophical positions, make different statistical assumptions, and can lead to incompatible empirical results (Edwards and Bagozzi 2000). As such, we agree with Welzel et al. (2021) that the distinction between reflective and formative indicators is of crucial importance for testing value theories and merits more



**Figure 1.** Graphical representation of reflective and formative measurement models.

scholarly attention. Yet, Welzel et al. (2021) neither provide the arguments why the value scales should follow a formative logic rather than a reflective one, nor do they present empirical evidence that the assumptions underlying the formative model are fulfilled. This is unfortunate, because – as Edwards and Bagozzi (2000:171) note – “If measures are specified as formative, their validity must still be established. It is bad practice to report a low reliability estimate, claim that one’s measures are formative, and do nothing more.”

*Reflective vs. formative indicators.* In the reflective model, indicators are conceived as manifestations of an underlying latent construct ( $\eta$  in Figure 1A). In line with classical test theory (Lord and Novick 2008), reflective indicators are assumed to contain a certain amount of measurement error ( $\delta$  in Figure 1A). The underlying logic of a reflective measurement is that the indicators are a subset of an infinite number of possible concrete manifestations of the latent variable. As such, reflective indicators are assumed to intercorrelate (as they are caused by the same latent phenomenon). Assessments of measurement quality hinge to a large extent on this internal consistency of the instrument. Furthermore, reflective indicators possess the characteristic of “useful redundancy” (Edwards 2011): they are interchangeable without altering the content of the construct (e.g., Brown 2015). Formative indicators, in contrast, are assumed to be determinants rather than consequences of the underlying latent variable (as depicted in Figure 1B). As each indicator contributes independently to the underlying concept, they can be seen as components of a latent variable. Contrary to reflective indicators, the formative model does not require that indicators are internally consistent. Instead, they might even be uncorrelated. At the same time, indicators are not regarded as interchangeable. Rather than being a random sample of possible items, the

indicators used to measure a formative construct should be a theoretically well-considered set of indicators that cover all relevant theoretical aspects of a concept and effectively define it.

Formative and reflective measurement models are neither equivalent nor nested; they are fundamentally different and rely on different underlying assumptions. Using either of them may lead to different associations with other theoretical constructs of interest, different mean scores, and different conclusions when used in comparative research (Bollen and Hoyle 2012). The decision to specify an indicator as formative or reflective is a question of the postulated causality between the construct and the indicator (Edwards and Bagozzi 2000). In the absence of experimental data, it is often difficult to determine which model is most appropriate on empirical grounds. Therefore, theoretical reasoning plays a crucial role, and a well-founded theoretical justification for the choice is crucial (Hardin 2017; Hempel & Oppenheim 1948).

In the context of comparative research, it is of crucial importance to stress that also *formative measurement requires that the constituent indicators are comparable*. For example, when comparing SES across countries – a concept often argued to be formative – we must still use schooling levels that have been harmonized if the comparison is to be sensible. The only difference with the reflective case in this regard is that reflective constructs have redundancy, making it much easier to establish (partial) invariance than in the formative case. The formative approach does not make comparability less important, but it does make it harder to examine.

*Formative or reflective measurement? The case of secular and emancipative values.* Thus, the decision to shift from reflective to formative measurement should not be taken lightly, as it requires convincing theoretical and empirical arguments. Unfortunately, such arguments are missing in the work of Welzel et al. (2021). Welzel et al. (2021) present a continuum between complementarity and similarity of indicators (Figure 1 in their paper), but they do not provide arguments on why their value dimensions should be considered as following a logic of complementarity (for a good example of how such an argumentation could look like, see Wuttke, Schimpf, and Schoen 2020).

In fact, when revisiting previous work of the authors, it becomes clear that there is an inconsistency over time with respect to the paradigm and methods they use to measure value dimensions. For example, Inglehart and Baker (2000) treat the items for measuring values as reflective indicators and factor-analytic tools (for reflective indicators) are used. Welzel (2010), on the one hand, claims to measure self-expression values by means of an index

following a formative logic, but on the other hand proceeds by justifying the measurement quality of the instrument using hierarchical factor analysis – a technique that assumes a reflective measurement. In the analyses reported in Dülmer, Welzel, and Inglehart (2015), the items are treated again as reflective indicators, as confirmatory factor analysis and multilevel confirmatory factor analysis techniques are used. And now Welzel and colleagues firmly argue that their value measurement is formative and thus “escapes the reflective logic of MGCFA” (2021:24).

This suggests that the authors’ rule for deciding whether values are formative or reflective appears to be this: choose whichever conception gives the result that the scale is “validated” and ignore the other. This allows Welzel et al. (2021:10) to admonish those who would question the comparability of their scale to stay on their “home turf”. But researchers can hardly be blamed for being confused regarding the turf claimed by the authors. Furthermore, if the scale truly is formative, then a large amount of work is still needed to establish that it is comparable.

### *The Argument of Restricted Realm of Applicability*

The second argument asserts that statistical methods of testing for measurement invariance have no use in empirical cross-country comparative analysis because of the closed-ended nature of the measures of the scales. They claim to have discovered a new result, namely that using closed-ended variables distorts results of MGCFA analysis, making non-invariance inevitable if there are groups in which the mean value of some variables are close to the endpoint of the scale (Welzel et al. 2021:12–17).

There are three problems with this argument. First, the result is not new, but an obvious consequence of the threshold model for categorical variables, which is due to Pearson (1900). Second, the problem as identified by the authors in their simulation can be easily solved by applying the categorical data model, a standard procedure in modern SEM. Third, in the authors’ setup, when treating the data as continuous – for example by summing the items – apparent differences in scores can indeed result even when there are no true differences in the underlying variable. In other words, when the data are (incorrectly) treated as continuous, there is indeed a violation of measurement invariance that can disturb the comparison.

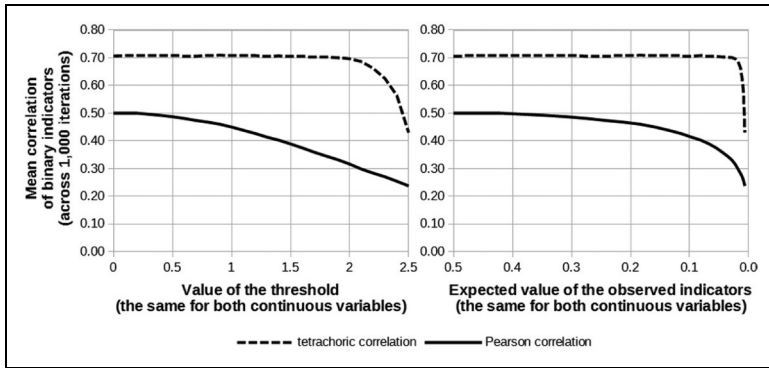
If indicators are closed-ended – either in the form of being “ordered-categorical” (like answers to the rating scale questions that can take only a few distinct values) or “censored” (like composite scores that can take quite many different values but only within some clearly defined limits), an

additional challenge arises for the correct estimation of measurement models. Methods that account for the categorical nature of data in the MGCFA family of models have been available for several decades (e.g., Bartholomew 1987; Flora and Curran 2004; Muthén 1984; Takane and de Leeuw 1987) and are commonly applied in current practice. Estimation approaches have been readily available in structural equation modeling software packages for more than a decade. Welzel and colleagues (2021) conveniently ignore these well-known solutions.

To provide a simple but persuasive example of how the aforementioned methods solve the problems caused by closed-ended or “censored” indicators, we present a small simulation. Let us assume that there are two standard-normally distributed variables, namely  $U_1$  and  $U_2$ , and that the correlation between them equals 0.7. Moreover, we assume that these variables cannot be observed directly, but that we observe binary indicators ( $X_1$  and  $X_2$ ) instead, which are created from  $U_1$  and  $U_2$  using a given threshold. For the sake of simplicity, we assume the threshold values are the same for both variables. We consider a threshold range between 0 and 2.5 with 26 conditions in which we increased the threshold consecutively in each condition by 0.1. In this way we simulate strong skewness and a mean of the observed indicators close to the endpoint, a situation for which Welzel et al. (2021:19) claim that it not possible to retrieve strong correlations. For each of the conditions, 1,000 replications were performed. Within each of the replications, 1,000 observations were sampled from a bivariate normal distribution for  $U_1$  and  $U_2$  and transformed into two binary indicators by cutting them at the threshold. Using these binary data, two estimates of the correlation between the underlying continuous variables  $U_1$  and  $U_2$  were computed: an ordinary Pearson product-moment correlation and a tetrachoric correlation.

Figure 2 compares these estimated correlations between the binary  $X$ -variables with the true 0.7 correlation in each condition. The two panels show average values (across replications performed in the same condition) of the estimated correlations as a function of the x-axis. The x-axis in the left panel displays different threshold values (ranging between 0 and 2.5). The x-axis in the right panel displays the thresholds’ corresponding expected values (i.e., the frequency of ones) for the binary indicators. Unsurprisingly, using ordinary Pearson product-moment correlations gives downward-biased estimates of the population correlation, even in the best scenario (i.e., when the threshold equals 0). Pearson correlations underestimate the true correlation between  $U_1$  and  $U_2$ ; as the threshold increases (and the distribution becomes more skewed), the downward bias becomes even larger. Tetrachoric correlations perform much better and can recover the correlation,





**Figure 2.** Consequences of using binary indicators of continuous variables on two different estimators of correlation (the real value of the correlation is 0.7; mean values of the correlation across 1,000 iterations are presented; 1,000 observations were used in each iteration).

almost irrespectively of the threshold and the corresponding expected value for the observed indicators. It is only when the threshold surpasses 2—corresponding roughly to a 0.975–0.025 split in the binary indicator—that the tetrachoric correlation starts to suffer from bias.<sup>1</sup> Clearly, using closed-ended indicators and mean disparity do not necessarily introduce downward bias in the correlation, provided that one applies the adequate statistical methods. This undercuts one of the core tenets of the argumentation of Welzel et al. (2021).

The authors’ argument can be summarized as follows. If we generate data according to the categorical thresholding model, and then analyze it using a linear model, the resulting misspecification will lead to a violation of the measurement invariance assumptions in the linear model. The authors conclude that this violation is spurious, because there is no such violation in the original thresholding model. However, this argument does not work, for the following reason. The linear model shows a violation, because the *linear* relationships of the items with their underlying construct are *indeed different* across countries. In consequence, if we simply treat the items as continuous – for example, by summing them – it is easy to create a situation in which the original “values” construct does not differ across the groups, but the sum scores do. In other words, the authors, endeavoring to show that invariance testing is futile, have unintentionally demonstrated its necessity.

### *The Argument of Nomological Linkages*

The third argument of Welzel et al. (2021:4) is that nomological linkages at the aggregate level offer an alternative guide for establishing measurement equivalence. In other words, the authors suggest that if a value scale, aggregated to the country level, correlates strongly with other country-level variables in a “theoretically expected” manner, this would be an indication of both the validity and the comparability of the scale. This argument resonates with the logic of construct validity (Cronbach and Meehl 1955). Construct validity certainly is a useful approach to measurement. However, the way in which Welzel et al. (2021) suggest applying this logic to cross-national comparisons of value patterns is problematic in several respects.

First, Welzel et al. (2021) depict the logic underlying internal consistency and external construct validation as being in opposition: In their view, the practice of invariance testing (rooted in the logic of internal validation) should be replaced by external validation through nomological networks. We disagree. Internal consistency and construct validity are two complementary approaches that are both necessary rather than sufficient conditions for establishing a good measurement. Furthermore, internal consistency has a crucial influence on the unbiasedness of external validation. Contrary to what Welzel et al. (2021; see also Welzel and Inglehart 2016:1,3,19) claim, Bollen (1989) shows that variations in random measurement error influence the correlation and prediction of criterion variables strongly; that is, random measurement error affects external validity.

Second, relying only on external relationships in nomological networks is particularly risky in the field of cross-national value research. In this field, differences in language and cultural context are quite common and can bias the external relationships that are estimated to validate the construct under scrutiny (see the previous point). Furthermore, aggregate-level analyses are further challenged by a small  $N$  at the country level and the fact that country-level variables tend to correlate strongly. We would like to further note that Welzel et al. (2021) artificially increase their sample size by aggregating to the country-year level rather than to the country level. This misrepresents the data structure and can lead to biased results, as Schmidt-Catran and Fairbrother (2016) have shown.

This specific situation of cross-cultural research makes aggregate-level analysis prone to spurious correlations. As a result, external validation through nomological linkages can be useful, but it is too weak a fundament to carry the full weight of scale validation. We argue thus that *both* construct and internal validity are needed rather than one or the other to meaningfully

analyze associations on the country level. Welzel and colleagues (2021), for example, maintain that “Across more than a hundred countries, including the biggest national populations in each global region, the EVI correlates in the theoretically expected ways with other, well-established markers (...). Many of these correlations are strikingly strong, often reaching an  $r$  of .80” (Welzel et al. 2021:6). They do not mention, however, that more appropriate methods could result in even higher correlations. For instance, Borgonovi and Pokropek (2020) showed that the trust in science scale measured in 144 countries by the Wellcome Global Monitor correlated higher with external variables (GDP and HDI) when accounting for measurement invariance using MGCFA with alignment optimization than when ignoring the need for measurement invariance. Similarly, Koc and Pokropek (2022) showed that political participation scales measured by the ESS were more strongly related to external criteria (GDP, polyarchy index, GINI) after accounting for measurement errors and controlling for measurement invariance by MGCFA alignment compared to when the need for measurement invariance is simply ignored.

Welzel et al. (2021) justify their decision to look exclusively at the aggregate level by referring to the “ontological primacy of the aggregate over the individual” (p. 11). While it is evident that culture influences the individual, this fact is not relevant for the problem at hand. The reason is quite straightforward: Welzel et al. (2021) use measures that are collected on the individual level and not on the cultural level. If one comes up with a way to measure cultural constructs of interest directly, that is, on the cultural level rather than on the individual level, then these constructs would not be computed as an aggregate of individual measures and would not be subject to psychometric rules that require reliability and validity of the measures on the individual level.

However, the research design of cross-cultural survey analysis, to which the authors themselves have contributed substantially, boils down to conducting measurements among individual respondents to make statements about cultural differences. As a result, individual-level psychometric requirements of reliability and validity necessarily apply to them (Billiet 2003; Chen 2008; De Beuckelaer 2005). If it turns out that the reliability and validity are poor, then aggregation makes no sense, because it is neither clear what the aggregate construct stands for nor whether, whatever it is, it is comparable across cultures (Davidov et al. 2014; see also Millsap 2011; Van de Vijver and Leung 1997; Van de Vijver and Poortinga 1997; Vandenberg 2002; Vandenberg and Lance 2000). In that sense, the claim that “the functioning of a multi-item index across groups depends in no way on the functioning

of its single items within groups” (Welzel et al. 2021:4–5) is manifestly incorrect (see the rapidly growing literature on invariance in multilevel settings: Jak, Oort, and Dolan 2013; 2014; Davidov et al. 2012; Ruelens, Meuleman, and Nicaise 2018).

To summarize, claiming that “theoretically expected” aggregate correlations are evidence of comparability is putting the cart before the horse. It is the other way around: we can only trust such correlations if measures are good enough and comparable enough. Millsap (2011) and Oberski (2014), for example, show in detail to what extent parameters of interest – such as correlations – are affected by bias when measurement invariance is not present.

## **Are Violations of Measurement Invariance Inconsequential? Some Illustrations**

The main issue at hand is the following: Can a violation of measurement invariance assumptions distort substantive findings substantially? Welzel et al. (2021:18) take the clear position that “non-invariance is obviously inconsequential for a construct’s cross-cultural functioning (...)”. Yet it is easy to show that neglecting measurement invariance can put comparative researchers on the wrong foot.

We illustrate this point by means of two brief examples.

First, Billiet (2013) showed how ignoring measurement invariance may lead to erroneous substantive conclusions. The European Social Survey (ESS) includes a scale measuring the level of religiosity. The scale consists of three questions inquiring about the level of religiosity and the frequency of religious practices. In one of the analyses of this scale, Turkey turned out to be the country with the lowest average level of religiosity in the ESS. A particularly low rate was observed among females. A closer look at the question items showed that the low average score of religiosity was a result of a low frequency of visits to religious services among Turkish women. This is of no surprise, because Islam is the dominant religion in Turkey. Indeed, the predominant form of Islam practiced in Turkey has different requirements for visiting religious services for males and females. If this difference is ignored during the construction of a scale measuring religious involvement, one would be misled in concluding that Turkish females are the least religious category in the ESS, although they are obviously not when looking at the other two indicators measuring religiosity in the ESS (Meuleman and Billiet 2018). Measurement invariance testing precisely

aims at identifying such items that behave in a different way in certain groups, and thus provides hints to applied researchers where they actually may identify noncomparability in their endeavor to improve the comparability of their scales.

For the second illustration, we use the work by Inglehart and Welzel (2005). According to data from the fourth round of the World Values Survey (WVS), no less than 99% of Vietnamese respondents appear to support military rule. This indicator feeds into their measurement of a pro-democratic civic culture and consequently suggests that support for democracy is extremely fragile in Vietnam. Inglehart and Welzel (2005: 266) interpret this striking finding as substantively relevant, pointing towards the specificity of national culture: “The very low percentage of ‘solid democrats’ in the case of Vietnam reflects a very high percentage of respondents expressing support for the army rule. In a country in which the army is a symbol of national liberation, these figures require a different interpretation.” What is even more striking is that in the subsequent wave of the WVS, the percentage of supporters of military rule has dropped to about one-third of the Vietnamese. Is it an indication of value change driven by generational replacement? Not at all.

Upon closer inspection, the score turns out to be the result of a simple translation error. In the fourth WVS round, Vietnamese respondents were asked to evaluate the “role of the military” rather than simply the “military” (Kurzman 2014). Such translation errors are hard to avoid in large-scale cross-national surveys, even with strict translation procedures. That is precisely the reason why we need measurement invariance tests that can reveal which items in which countries behave differently, so that substantive researchers can examine such items more closely rather than being misled by spurious results. External validation through nomological linkages cannot fulfill this role entirely, because the presence of this translation error has the power to bias the estimated relationship with external concepts. The work of Inglehart and Welzel (2005: 266) illustrates this point clearly (and thus contradicts the statements made by Welzel et al. 2021): “Figure 11.2b shows the impact of mass preferences for democracy versus autocracy (the percentage of ‘solid democrats’) on effective democracy, controlling for the strength of self-expression values. There is a weak relationship that only exists because of one single leverage case: Vietnam.”

## Final Remarks

In sum, the criticism of Welzel et al. (2021) on the practice of testing for measurement invariance is largely based on a misrepresentation of the field

and incorrect arguments. With this response, we want to direct readers' attention to a large body of literature which either formally proves the points we refer to or explains in much more detail why measurement invariance testing is crucial for a meaningful cross-group comparison. We hope that we were able to clear up at least some of the arguments raised against the application of measurement invariance testing. The appendix provides a more detailed discussion of some of the arguments of Welzel et al. (2021), including a simulation study. Furthermore, this contribution mentions different methods available to examine measurement invariance that do not require exact invariance but are more lenient or make fewer assumptions on the data. This underlines the current state of knowledge that while exact measurement invariance across groups is not necessary in all cases, some level of invariance is crucial for conducting meaningful comparative survey research.

Ultimately, researchers should heed the comparability of measures not because of some imagined methodological authority, but because if they do not, claims might not be veracious. With this in mind, we hope that methodologists and applied researchers can agree to keep sight of the goal: discovering true, rather than merely apparent, facts about cross-cultural similarities and differences.

### **Declaration of Conflicting Interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### **Funding**


The author(s) received no financial support for the research, authorship, and/or publication of this article.


### **Author's Note**


Daniel L. Oberski is also affiliated to Department of Data Science, University Medical Center Utrecht. Peter Schmidt is also affiliated to Department of Psychosomatic Medicine, University of Mainz.

### **ORCID iDs**

Bart Meuleman  <https://orcid.org/0000-0002-0384-5995>

Tomasz Żółtak  <https://orcid.org/0000-0003-1354-4472>

Artur Pokropek  <https://orcid.org/0000-0002-5899-2917>

Eldad Davidov  <https://orcid.org/0000-0002-3396-969X>

## Supplemental material

Supplemental material for this article is available online.

## Note

1. By increasing the sample size (1000 in this simulation), the range of threshold values for which tetrachoric correlation performs well could be increased even further.

## References

- Alemán, José and Dwayne Woods. 2016. "Value Orientations from the World Values Survey." *Comparative Political Studies* 49:1039-67.
- Asparouhov, Tihomir and Bengt O. Muthén. 2014. "Multiple-group factor analysis alignment." *Structural Equation Modeling: A Multidisciplinary Journal* 21(4):495-508.
- Bartholomew, David J. 1987. *Latent Variable Models and Factor Analysis*. New York: Oxford University Press.
- Billiet, Jaak. 2003 "Cross-Cultural Equivalence with Structural Equation Modeling." Pp. 247-64 in *Cross-Cultural Survey Methods*, edited by Janet A. Harkness, Fons J. R. Van de Vijver, and Peter P. Mohler. New York: John Wiley.
- Billiet, Jaak. 2013 "Quantitative Methods With Survey Data in Comparative Research." Pp. 264-300 in *A Handbook of Comparative Social Policy*, 2nd ed., edited by Patricia Kennett. Cheltenham, UK: Edward Elgar.
- Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: John Wiley & Sons.
- Bollen, Kenneth. A. and Rick H. Hoyle. 2012 "Latent Variables in Structural Equation Modeling." Pp. 56-67 in *Handbook of Structural Equation Modeling*, edited by Rick H. Hoyle. New York: Guilford Press.
- Borgonovi, Francesca and Artur Pokropek. 2020. "Can We Rely on Trust in Science to Beat the COVID-19 Pandemic?" *PsyArXiv*. May 21. doi: 10.31234/osf.io/yq287
- Brown, Timothy A. 2015. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press.
- Byrne, Barbara M., Richard J. Shavelson, and Bengt O. Muthén. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105:456-66.
- Chen, Fang Fang. 2008. "What Happens if We Compare Chopsticks with Forks? The Impact of Making Inappropriate Comparisons in Cross-Cultural Research." *Journal of Personality & Social Psychology* 95(5):1005-18.

- Cronbach, Lee J. and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52(4):281-302.
- Davidov, Eldad, Hermann Dülmer, Elmar Schlueter, Peter Schmidt, and Bart Meuleman. 2012. "Using a Multilevel Structural Equation Modeling Approach to Explain Cross-Cultural Measurement Noninvariance." *Journal of Cross-Cultural Psychology* 43(4):558-75.
- Davidov, Eldad, Bart Meuleman, Jan Cieciuch, Peter Schmidt, and Jaak Billiet. 2014. "Measurement Equivalence in Cross-National Research." *Annual Review of Sociology* 40:55-75.
- Davidov, Eldad, Bengt Muthén, and Peter Schmidt. 2018. "Measurement Invariance in Cross-National Studies: Challenging Traditional Approaches and Evaluating New Ones." *Sociological Methods & Research* 47(4):631-36.
- De Beuckelaer, Alain. 2005. "Measurement Invariance Issues in International Management Research." PhD dissertation. Hasselt University, Diepenbeek, Belgium.
- Dülmer, Hermann, Ronald Inglehart, and Christian Welzel. 2015. "Testing the Revised Theory of Modernization: Measurement and Explanatory Aspects." *World Values Research* 8(2):68-100.
- Edwards, Jeffrey R. 2011. "The Fallacy of Formative Measurement." *Organizational Research Methods* 14(2):370-88.
- Edwards, Jeffrey. R. and Richard P. Bagozzi. 2000. "On the Nature and Direction of Relationships Between Constructs and Measures." *Psychological Methods* 5(2):155-74.
- Flora, David B. and Patrick J. Curran. 2004. "An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data." *Psychological Methods* 9(4):466-91.
- Hardin, Andrew. 2017. "A Call for Theory to Support the Use of Causal-Formative Indicators: A Commentary on Bollen and Diamantopoulos (2017)." *Psychological Methods* 22(3):597-604.
- Hempel, Carl G. and Paul Oppenheim. 1948. "Studies in the Logic of Explanation." *Philosophy of Science* 15(2):135-75.
- Holland, Paul W. and Howard Wainer, eds. 2012. *Differential Item Functioning*. New York: Routledge.
- Inglehart, Ronald and Wayne E. Baker. 2000. "Modernization, Cultural Change, and the Persistence of Traditional Values." *American Sociological Review* 65:19-51.
- Inglehart, Ronald and Christian Welzel. 2005. *Modernization, Cultural Change, and Democracy*. New York: Cambridge University Press.
- Jak, Suzanne, Frans J. Oort, and Conor V. Dolan. 2013. "A Test for Cluster Bias: Detecting Violations of Measurement Invariance Across Clusters in Multilevel Data." *Structural Equation Modeling: A Multidisciplinary Journal* 20(2):265-82.



- Jöreskog, Karl G. 1971. "Simultaneous Factor Analysis in Several Populations." *Psychometrika* 36(4):409-26.
- Koc, Piotr and Artur Pokropek. 2022. "Accounting for Cross-Country-Cross-Time Variations in Measurement Invariance Testing. A Case of Political Participation." *Survey Research Methods* 16(1): Forthcoming.
- Kurzman, Charles. 2014. "World Values Lost in Translation." *Washington Post*, September 2. Retrieved from <https://www.washingtonpost.com/news/monkey-cage/wp/2014/09/02/world-values-lost-in-translation/>.
- Lord, Frederic M. and Melvin R. Novick. 2008. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.
- Mellenbergh, Gideon J. 1989. "Item Bias and Item Response Theory." *International Journal of Educational Research* 13(2):127-43.
- Meredith, William and Roger E. Millsap. 1992. "On the Misuse of Manifest Variables in the Detection of Measurement Bias." *Psychometrika* 57:289-311.
- Meuleman, Bart and Jaak Billiet. 2018 "Religious Involvement: Its Relation to Values and Social Attitudes." Pp. 181-214 in *Cross-Cultural Analysis*, edited by Eldad Davidov, Peter Schmidt, Jaak Billiet, and Bart Meuleman. New York: Routledge.
- Millsap, Roger E. 2011. *Statistical Approaches to Measurement Invariance*. New York: Taylor and Francis Group.
- Muthén, Bengt O. 1984. "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators." *Psychometrika* 49(1):115-32.
- Oberski, Daniel L. 2014. "Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models." *Political Analysis* 22(1):45-60.
- Pearson, Karl. 1900. "Mathematical Contributions to the Theory of Evolution. —VII. On the Correlation of Characters Not Quantitatively Measurable." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 195:1-47.
- Ruelens, Anna, Bart Meuleman, and Ides Nicaise. 2018. "Examining Measurement Isomorphism of Multilevel Constructs: The Case of Political Trust." *Social Indicators Research* 140:907-27.
- Schmidt-Catran, Alexander W. and Malcolm Fairbrother. 2016. "The Random Effects in Multilevel Models: Getting Them Wrong and Getting Them Right." *European Sociological Review* 32(1):23-38.
- Sokolov, Boris. 2018. "The Index of Emancipative Values." *American Political Science Review* 112:395-408.
- Steenkamp, Jan-Benedict E. M. and Hans Baumgartner. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25(1):78-90.

- Takane, Yoshio and Jan de Leeuw. 1987. "On the Relationship Between Item Response Theory and Factor Analysis of Discredited Variables." *Psychometrika* 52(3):393-408.
- Van de Vijver, Fons J. R. and Kwok Leung. 1997. *Methods and Data Analysis for Cross-Cultural Research*. London: Sage.
- Van de Vijver, Fons J. R. and Ype H. Poortinga. 1997. "Towards an Integrated Analysis of Bias in Cross-Cultural Assessment." *European Journal of Psychological Assessment* 13(1):29-37.
- Vandenberg, Robert J. 2002. "Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures." *Organizational Research Methods* 5(2):139-58.
- Vandenberg, Robert J. and Charles E. Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3(1):4-70.
- Welzel, Christian. 2010. "How Selfish are Self-Expression Values? A Civicness Test." *Journal of Cross-Cultural Psychology* 41(2):152-74.
- Welzel, Christian, Lennart Brunkert, Ronald F. Inglehart, and Stefan Kruse. 2019. "Measurement Equivalence? A Tale of False Obsessions and a Cure." *World Values Research* 11(3):54-84.
- Welzel, Christian, Lennart Brunkert, Stefan Kruse, and Ronald F. Inglehart. 2021. "Non-invariance? An Overstated Problem With Misconceived Causes." *Sociological Methods & Research*. doi: 10.1177/0049124121995521
- Welzel, Christian and Ronald F. Inglehart. 2016. "Misconceptions of Measurement Equivalence: Time for a Paradigm Shift." *Comparative Political Studies* 49(8): 1068-94.
- Wuttke, Alexander, Christian Schimpf, and Harald Schoen. 2020. "When the Whole is Greater Than the Sum of its Parts: On the Conceptualization and Measurement of Populist Attitudes and Other Multidimensional Constructs." *American Political Science Review* 114(2):356-74.

## Author Biographies

**Bart Meuleman** is a full professor at the Centre for Sociological Research (CeSO) at the University of Leuven (Belgium). His main research interests are cross-cultural survey methodology and cross-national comparisons of value and attitude patterns.

**Tomasz Żółtak** is a post-doctoral researcher at the Institute of Philosophy and Sociology, Polish Academy of sciences, where he is involved in research on survey methodology and psychometry.

**Artur Pokropek** is a Professor at the Institute of Philosophy and Sociology of the Polish Academy of Sciences and faculty member of the Educational Research Institute in Warsaw. His main areas of research interests are statistics, research methods, evaluation, psychometrics and machine learning.

**Eldad Davidov** is professor at the Institute for Sociology and Social Psychology at the University of Cologne, Germany, and the Department of Sociology at the University of Zurich, Switzerland, and co-director of the University of Zurich Research Priority Program “Social Networks”. He was president of the European Survey Research Association (ESRA) between 2015 and 2017. His research interests are applications of structural equation modeling to survey data, especially in cross-cultural research. Applications include human values and attitudes toward immigrants and other minorities

**Bengt Muthén** obtained his Ph.D. in Statistics at the University of Uppsala, Sweden and is Professor Emeritus at UCLA. He was the 1988-89 President of the Psychometric Society and the 2011 recipient of the Psychometric Society’s Lifetime Achievement Award. He was the 2017 recipient of the Sells Award for Outstanding Career Contributions to Multivariate Experimental Psychology from the Society of Multivariate Experimental Psychology. He has published extensively on latent variable modeling and many of his procedures are implemented in the Mplus software.

**Daniel L. Oberski** is full professor of health and social data science at the department of methodology & statistics, Utrecht University, where he leads the Human Data Science group; and at the University Medical Center Utrecht, where he leads the department of Data Science.

**Jaak Billiet** is Emeritus Professor of Social Methodology, Centre of Sociological Research, University of Leuven. He combines methodological research with substantial longitudinal and comparative research on ethnocentrism, political attitudes and religious orientations.

**Peter Schmidt** is Professor Emeritus at the Department of Political Science and Member at the Centre for International Development and Environment(ZEU) at the University of Giessen and Research Fellow and Principal Investigator at the Department of Psychosomatics at the University of Mainz.