Survey Research Methods (2020)
Vol. 14, No. 4, pp. 345-349
doi:10.18148/srm/2020.v14i4.7655

ISSN 1864-3361
http://www.surveymethods.org

Measurement Invariance: Testing for It and Explaining Why It is Absent

Katharina Meitinger
Faculty for Social and Behavioural Sciences
Utrecht University
The Netherlands

Eldad Davidov
Faculty of Management, Economics and Social Sciences
University of Cologne, Germany and
URPP "Social Networks" and Department of Sociology
University of Zurich, Switzerland and
The Minerva Center on Intersectionality in Aging
University of Haifa, Israel

Peter Schmidt
Centre for international Development and Environmental Research (ZEU)
University of Giessen, Germany

Michael Braun
GESIS—Leibniz Institute for the Social Sciences
Mannheim, Germany

There has been a significant increase in cross-national and longitudinal data production in social science research in recent decades. Before drawing substantive conclusions based on cross-national and longitudinal survey data, researchers need to assess whether the constructs are measured in the same way across countries and time-points. If cross-national data are not tested for comparability, researchers risk confusing methodological artefacts as "real" substantive differences across countries. However, researchers often find it particularly difficult to establish the highest level of measurement invariance, that is, exact scalar invariance. When measurement invariance is rejected, it is crucial to understand why this was the case and to address its absence with approaches, such as alignment optimization or Bayesian structural equation modelling.

Keywords: measurement equivalence; comparability; bias; approximate measurement invariance; alignment, BSEM

There has been a significant increase in cross-national and longitudinal data production in social science research in recent decades (Johnson, Pennell, Stoop, & Dorer, 2019). Before drawing substantive conclusions based on cross-national and longitudinal survey data, it is necessary to assess whether the constructs are measured in the same way across countries and time-points (Cieciuch, Davidov, Schmidt, & Algesheimer, 2019). If cross-national data are not tested for comparability, researchers risk confusing methodological artifacts as "real" substantive differences across countries (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Steenkamp & Baumgartner, 1998); for an overview over different approaches to assess comparability, see Braun and Johnson (2010).

Measurement invariance (MI) tests are an increasingly popular way of assessing the cross-national and longitudinal comparability of survey data (see for example Davidov, 2008; Davidov, Cieciuch, & Schmidt, 2018; Weber, 2011). The so-called exact MI tests that use multigroup confirmatory factor analysis (MGCFA; see Jöreskog, 1971) are an approach to assess comparability of survey data measures across groups. In general, researchers distinguish at least three levels of comparability when applying MGCFA: configural, metric, and scalar MI, which provide insights into whether the constructs, the coefficients, and the latent means of a construct can be compared with confidence across units of analysis (Meredith, 1993; Millsap, 2011; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). If configural and metric invariance is established, the latent concept can be meaningfully discussed across units of analysis and it is possible to compare structural relationships, such as unstandardized regression coefficients or covariances (Meredith, 1993; Millsap, 2011; Steenkamp & Baumgartner, 1998). Finally, achieving scalar invariance is a precondition for comparing mean values across groups or time points (Davidov et al., 2014).

However, researchers often find it particularly difficult to establish the highest level of measurement invariance, that is, scalar invariance (Davidov, Dülmer, Schlüter, Schmidt, & Meuleman, 2012). As a consequence, it is often the case

Contact information: Katharina Meitinger. Padualaan 14, 3584 CH Utrecht, The Netherlands (E-mail: k.m.meitinger@uu.nl)

that the latent means of constructs cannot be meaningfully compared (Meredith, 1993; Millsap, 2011). This also has consequences for further analyses because approaches such as multilevel analyses (Hox, Moerbeek, & van de Schoot, 2017) or a ranking of countries according to the constructs' mean values cannot be performed with confidence anymore (for robustness studies on this topic, see Meuleman, 2012; Oberski, 2014).

The finding that scalar MI is rarely achieved is not too surprising. Indeed, MI testing in the context of MGCFA examines the exact equality of parameters (factor loadings and item intercepts) across units of analysis and is thus very demanding. However, it might be possible that the parameter differences between units of analysis are actually negligible and do not matter for the substantive interpretation of the parameters of interest's differences across groups. Such small and negligible measurement parameter differences cannot be allowed in MGCFA models. Since higher levels of measurement invariance can rarely be established in cross-national studies, especially if one compares a large number of time points, countries and constructs simultaneously, the exact approach has been criticized as being too strict in recent years (e.g., van de Schoot et al., 2013; Zercher, Schmidt, Cieciuch, & Davidov, 2015).

Questionnaire development and data collection in a multicultural, multinational and/or multilinguistic context (3MC) is very challenging and adds several "layers of complexity" (Lynn, Japec, & Lyberg, 2006) to the generation of data (see also Johnson et al., 2019). Differences in parameters of interest across groups in 3MC data may be methodological artifacts and not substantive results for a variety of reasons, also called biases. Bias can be seen as "nuisance factors that jeopardize the validity of instruments applied in different cultures" (He & van de Vijver, 2012, p. 3; van de Vijver, 2018). Construct bias means that the measured construct differs across cultures (van de Vijver & Poortinga, 1997), whereas distorting effects through specific methods and the context of the measurement can create a method bias, for example, due to differences in sampling procedures or modes of data collection (He & van de Vijver, 2012). Additionally, poor item translation, ambiguous source items, inapplicability of item contents or connotations associated with the item wording in some countries but not in others can affect the comparability of items and create an item bias (He & van de Vijver, 2012; van de Vijver & Leung, 2011).

When MI is rejected, it is crucial to understand why this was the case. Was the method of MI testing too strict? Was it because the content of the construct differed across groups? Did respondents attribute different meanings and associations with the questions (construct and item bias)? Or was the failure to reach MI a result of other sources of measurement error (e.g., method bias)? Disentangling the sources of bias is vital to develop strategies to address a lack of MI and improve the development of measurement instruments.

In recent years, statistical solutions have been proposed to answer the question of whether MI testing with MGCFA is too strict. Recent approaches such as Bayesian structural equation modeling (BSEM; Muthén & Asparouhov, 2012; van de Schoot et al., 2013) or alignment (Asparouhov & Muthén, 2014) propose to relax certain requirements when testing for MI. These approaches are promising because they allow for small parameter differences across countries which allows for more leeway during the assessment of measurement invariance (Seddig & Leitgöb, 2018; van de Schoot et al., 2013). As such, they may often suggest that approximate measurement invariance is given while more traditional, stricter approaches indicate that it is absent. This means that they may allow researchers to perform meaningful comparative analyses more frequently if the parameter difference across units of analysis is small enough.

When one realizes that even approximate invariance is not given, the possibility exists to identify the factors that reduce comparability. Several approaches have been proposed in this context. Quantitative approaches—such as the multiple indicators multiple causes model (MIMIC; Davidov et al., 2014; Marsh et al., 2018) and multilevel structural equation models (MLSEMs Davidov et al., 2018; Davidov et al., 2012)—view the lack of measurement invariance as a source of information on cross-group differences, try to explain the individual, social or historical sources of measurement nonequivalence (Davidov et al., 2014; Jak, Oort, & Dolan, 2013) and thus aim to substantively explain the sources of non-invariance. Unfortunately, some of these approaches require a relatively large number of groups to be applied, which is often not given in cross-cultural survey research. At the same time, there is an increasing awareness of the potential of mixed-method approaches to explain instances of measurement non-invariance. These methods combine measurement invariance tests with different qualitative or quantitative approaches (e.g., Benítez & Padilla, 2014; Latcheva, 2011; Meitinger, 2017) that try to identify sources of non-invariance. A third strategy in the literature has been to assess the relative impact of different sources of measurement error (such as mode, response scales or response styles) on MI (e.g., Hox, de Leeuw, & Zijlmans, 2015).

The 13 contributions for the session on "Measurement Invariance: Testing for It and Explaining Why It is Absent" (at the ESRA Conference in Lisbon, 2017) revealed a variety of innovative approaches to address the challenge of measurement non-invariance. They suggest that this topic is vividly studied and that there are many innovative and underexplored roads in the analysis of measurement invariance in comparative survey data and in the explanations of its absence.

Four contributions are included in this special issue (see Table 1). The first two are related to addressing the absence of exact scalar measurement invariance. The paper

Table 1
An overview of the contributions by authors' names, title, methodological focus, method, scale and the dataset used

Author(s)	Title	Methodological focus	Method	Scale	Data
Seddig, Maskileyson, and Davidov (2020)	The comparability of measures in the ageism module of the fourth round of the European Social Survey, 2008-2009	How to deal with missing exact scalar measurement invariance?	MGCFA[a] and alignment optimization	(1) Competence and warmth; (2) Experience of age discrimination.	European Social Survey 4th round (2008-2009): 29 European countries
Lytkina (2020)	Revisiting the Middleton Alienation Scale: In search of a cross-culturally valid instrument	How to deal with missing exact scalar measurement invariance?	MGCFA;[a] BSEM[b] and discriminant validity tests	(1) Anomie; (2) Alienation.	World Value Study 6th round (2011): Russia and Kazakhstan. Euromodule (1999-2002): Slovenia, Germany, Hungary, Spain, Switzerland, Austria, Turkey, and South Korea
Roberts, Sarrasin, and Stähli (2020)	Investigating the relative impact of different sources of measurement non-equivalence in comparative surveys: An illustration with scale format, data collection mode and cross-national variations	Why are data (not) comparable? Assessed sources of noninvariance: (1) Scale; (2) Mode; (3) Within and between country cultural variation.	Full and partial MGCFA[a]	(1) Evaluative wellbeing; (2) Emotional wellbeing.	European Social Survey 3rd round (2006): Switzerland (German and French speaking), Germany and France
Lee, Vasquez, Ryan, and Smith (2020)	Measurement equivalence of subjective well-being measures under the presence of acquiescent response style for the racially and ethnically diverse older population in the United States	Why are data (not) comparable? (1) Within country cultural variation; (2) Acquiescence response style.	Full and partial MGCFA and validity tests	(1) Satisfaction with life; (2) Positive affect; (3) Purpose in life.	Health and Retirement Study (2010): non-Hispanic Whites, Hispanics interviewed in English and Hispanics interviewed in Spanish in the United States

[a] Multi-group confirmatory factor analysis [b] BSEM: Bayesian structural equation modeling,

by Seddig, Maskileyson, and Davidov investigates measures of the ageism module in the European Social Survey and complements MI testing in the context of MGCFA with the more liberal alignment optimization procedure. The paper by Lytkina assesses the cross-national comparability of the constructs of alienation and anomie with the stricter MGCFA approach and complements it with approximate MI testing using Bayesian SEM. The last two contributions address the question of why data are (not) comparable. Both papers study the comparability of well-being measures. The paper by Roberts, Sarrasin, and Stähli evaluates the influence of scale format, data collection mode, and cultural variation on MI. Lee, Vasquez, Ryan, and Smith study the influence of ethnicity, language, and acquiescence response style on MI.

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508.

Benítez, I., & Padilla, J. L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research*, *8*(1), 52–68.

Braun, M., & Johnson, T. P. (2010). An illustrative review of techniques for detecting inequivalences. In J. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. Mohler, ... T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 375–393). Hoboken, NJ: Wiley-Blackwell.

Cieciuch, J., Davidov, E., Schmidt, P., & Algesheimer, R. (2019). How to obtain comparable measures for cross-national comparisons. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *71*, 157–186.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European Social Survey. *Survey Research Methods*, *2*(1), 33–46.

Davidov, E., Cieciuch, J., & Schmidt, P. (2018). The cross-country measurement comparability in the immigration module of the European Social Survey 2014-15. *Survey Research Methods*, *12*(1), 15–27.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*(4), 558–575.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75.

He, J., & van de Vijver, F. J. R. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, *2*(2). doi:10.9707/2307-0919.1111

Hox, J. J., de Leeuw, E. D., & Zijlmans, E. A. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in psychology*, *6*, 87.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. New York: Routledge.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling*, *20*(2), 265–282.

Johnson, T. P., Pennell, B.-E., Stoop, I., & Dorer, B. ((2019). *Advances in comparative survey methods: Multinational, multiregional and multicultural contexts (3MC)*. Hoboken, New Jersey: John Wiley & Sons.

Jöreskog, K. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–26.

Latcheva, R. (2011). Cognitive interviewing and factor-analytic techniques: A mixed method approach to validity of survey items measuring national identity. *Quality & Quantity*, *45*(6), 1175–1199.

Lee, S., Vasquez, E., Ryan, L., & Smith, J. (2020). Measurement equivalence of subjective well-being measures in the Health and Retirement Study with the racially and ethnically diverse older population in the United States. *Survey Research Methods*, *14*, xxx. doi:10.18148/srm/2020.v14i4.7413

Lynn, P., Japec, L., & Lyberg, L. (2006). What's so special about cross-national surveys? In J. Harkness (Ed.), *Conducting cross-national and cross-cultural surveys* (pp. 7–21). Mannheim: Zuma.

Lytkina, E. (2020). The Middleton alienation scale: Testing for measurement invariance. *Survey Research Methods*, *14*, xxx. doi:10.18148/srm/2020.v14i4.7421

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

Meitinger, K. (2017). Necessary but insufficient. why measurement invariance tests need online probing as a complementary tool. *Public Opinion Quarterly*, *81*(2), 447–472.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Meuleman, B. (2012). When are item intercept differences substantively relevant in measurement invariance testing. In S. Salzborn, E. Davidov, & J. Reinecke (Eds.),

Methods, theories, and empirical applications in the social sciences: Festschrift for Peter Schmidt (pp. 97–104). Wiesbaden: Springer.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York: Taylor & Francis.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335.

Oberski, D. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60.

Roberts, C., Sarrasin, O., & Stähli, M. (2020). The relative impact of different sources of measurement non-equivalence in comparative surveys: Examples of item, method and construct bias. *Survey Research Methods*, *14*, xxx. doi:10.18148/srm/2020.v14i4.7416

Seddig, D., & Leitgöb, H. (2018). Approximate measurement invariance and longitudinal confirmatory factor analysis: Concept and application with panel data. *Survey Research Methods*, *12*(1), 29–41.

Seddig, D., Maskileyson, D., & Davidov, E. (2020). The comparability of measures in the ageism module of the fourth round of the European Social Survey, 2008-2009. *Survey Research Methods*, *14*, xxx. doi:10 . 18148/srm/2020.v14i4.7369

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research*, *25*(1), 78–90.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J. J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology*, *4*, 173–187.

van de Vijver, F. J. R. (2018). Capturing bias in structural equation modeling. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross cultural analysis: Methods and applications* (pp. 3–44). London: Routledge.

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Culture and psychology. cross-cultural research methods in psychology* (pp. 17–45). Cambridge: Cambridge University Press.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*(1), 29–37.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4–70.

Weber, W. (2011). Testing for measurement equivalence of individuals' left-right orientation. *Survey Research Methods*, *5*(1), 1–10.

Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: Exact versus approximate measurement equivalence. *Frontiers in psychology.*, *6*(733), 1–11.