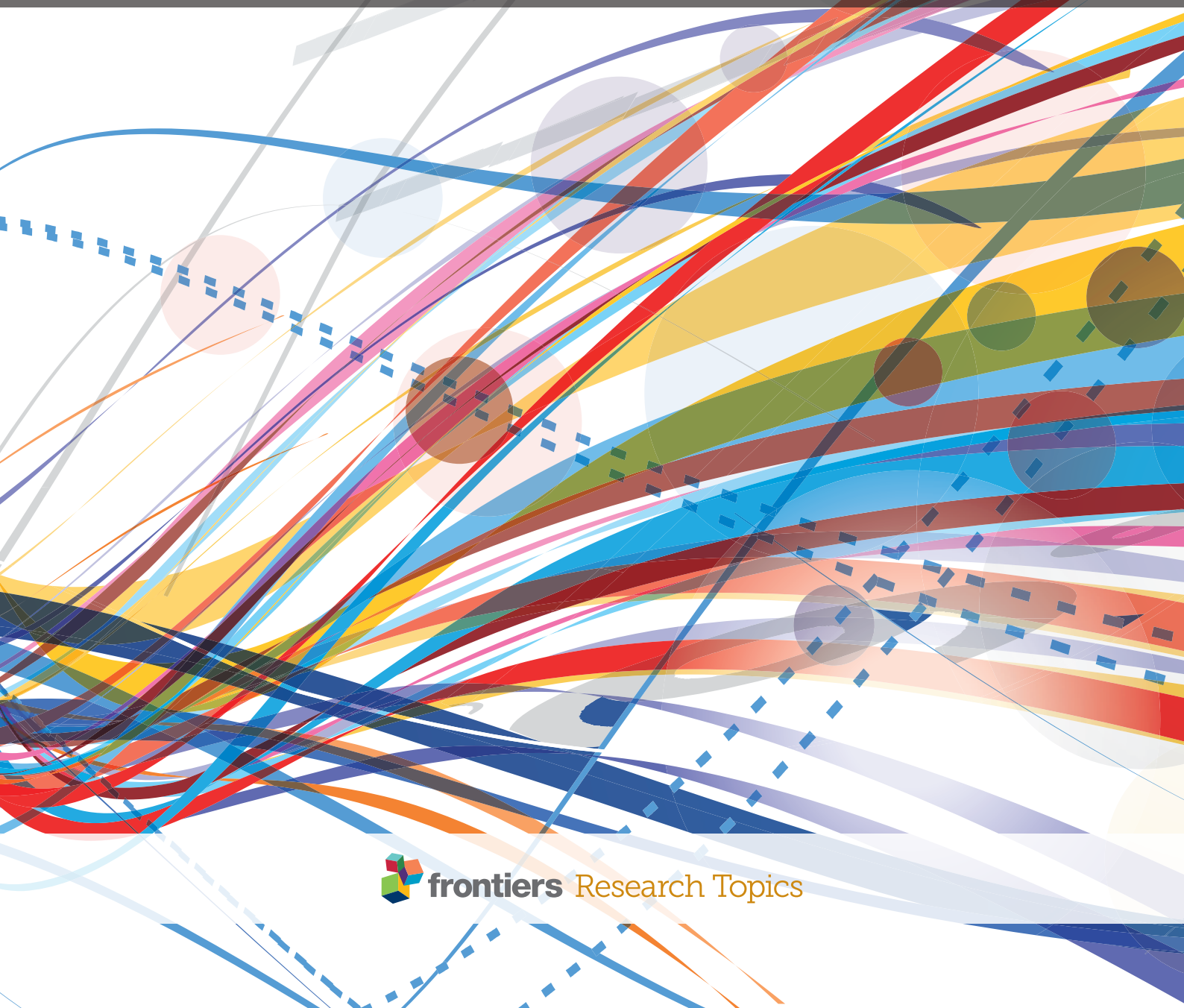
MEASUREMENT INVARIANCE

EDITED BY: Rens van de Schoot, Peter Schmidt and Alain De Beuckelaer
PUBLISHED IN: Frontiers in Psychology



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88919-650-0
DOI 10.3389/978-2-88919-650-0

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

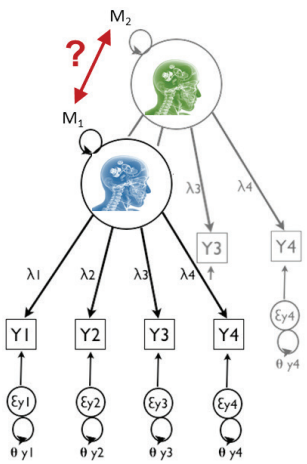
Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

MEASUREMENT INVARIANCE

Topic Editors:
Rens van de Schoot, Utrecht University, Netherlands and North-West University, South Africa
Peter Schmidt, University of Giessen, Germany and Research University Higher School of Economics, Russia
Alain De Beuckelaer, Institute for Management Research, Radboud University Nijmegen, Netherlands, Renmin University of China, China and Ghent University, Belgium



If latent factor means (M1 vs. M2) are to be meaningfully compared across groups or over time, the measurement structures including the latent factor and their survey items should be stable, that is 'invariant', across groups or over time. Image by Rens Van De Schoot.

Multi-item surveys are frequently used to study scores on latent factors, like human values, attitudes and behavior. Such studies often include a comparison, between specific groups of individuals, either at one or multiple points in time. If such latent factor means and relations between constructs are to be meaningfully compared, the measurement structures including the latent factor and their survey items should be stable across groups and/or over time, that is 'invariant'. Recent developments in statistics have provided new analytical tools for assessing measurement invariance (MI). The aim of this special issue is to provide a forum for a discussion of MI , covering some crucial 'themes': (1) ways to assess and deal with measurement non-invariance; (2) Bayesian and IRT methods employing the concept of approximate measurement invariance; and (3) new or adjusted approaches for testing MI to fit increasingly complex statistical models and specific characteristics of survey data.

The special issue started with a kick-off meeting where all potential contributors shared ideas on potential papers. This expert workshop was organized at Utrecht University in The Netherlands and was funded by the Netherlands Organization for Scientific Research (NWO-VENI-451-11-008). After the kick-off meeting the authors submitted their papers, all of which were reviewed by experts in the field. The papers in the eBook are listed in alphabetical order, but in the editorial the papers are introduced thematically.

Although it is impossible to cover all areas of relevant research in the field of MI, papers in this eBook provide insight on important aspects of measurement invariance. We hope that the discussions included in this special issue will stimulate further research on MI and facilitate further discussions to support the understanding of the role of MI in multi-item surveys.

Citation: van de Schoot, R., Schmidt, P., De Beuckelaer, A., eds. (2015). Measurement Invariance. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-650-0

Table of Contents

frontiers in Psychology

EDITORIAL
published: 28 July 2015
doi: 10.3389/fpsyg.2015.01064

Editorial: Measurement Invariance

Rens Van De Schoot[1,2]*, *Peter Schmidt*[3,4], *Alain De Beuckelaer*[5,6,7], *Kimberley Lek*[1] and *Marielle Zondervan-Zwijnenburg*[1]

[1] Department of Methods and Statistics, Utrecht University, Utrecht, Netherlands, [2] Optentia Research Program, Faculty of Humanities, North-West University, Vanderbijlpark, South Africa, [3] Department of Political Sciences, University of Giessen, Giessen, Germany, [4] International Laboratory of Social-Cultural Research, Research University Higher School of Economics, Moscow, Russia, [5] Nijmegen School of Management, Institute for Management Research, Radboud University Nijmegen, Nijmegen, Netherlands, [6] School of Sociology and Population Studies, Renmin University of China, Beijing, China, [7] Department of Personnel Management and Work and Organizational Psychology, Ghent University, Ghent, Belgium

Keywords: measurement invariance, confirmatory factor analysis, Bayesian models, questionnaires, approximate measurement invariance, partial measurement invariance

Multi-item surveys are frequently used to study scores on latent factors, like human values, attitudes, and behavior. Such studies often include a comparison, between specific groups of individuals or residents of different countries, either at one or multiple points in time (i.e., a cross-sectional or a longitudinal comparison or both). If latent factor means are to be meaningfully compared, the measurement structures of the latent factor and their survey items should be stable, that is "invariant." As proposed by Mellenbergh (1989), "measurement invariance" (MI) requires that the association between the items (or test scores) and the latent factors (or latent traits) of individuals should not depend on group membership or measurement occasion (i.e., time). In other words, if item scores are (approximately) multivariate normally distributed, conditional on the latent factor scores, the expected values, the covariances between items, and the unexplained variance unrelated to the latent factors should be equal across groups.

Many studies examining MI of survey scales have shown that the MI assumption is very hard to meet. In particular, strict forms of MI rarely hold. With "strict" we refer to a situation in which measurement parameters are exactly the same across groups or measurement occasions, that is an enforcement of zero tolerance with respect to deviations between groups or measurement occasions. Often, researchers just ignore MI issues and compare latent factor means across groups or measurement occasions even though the psychometric basis for such a practice does not hold. However, when a strict form of MI is not established and one must conclude that respondents attach different meanings to survey items, this makes it impossible to make valid comparisons between latent factor means. As such, the potential bias caused by measurement non-invariance obstructs the comparison of latent factor means (if strict MI does not hold) or regression coefficients (if less strict forms of MI do not hold).

Traditionally, MI is tested for in a multiple group confirmatory factor analysis (MGCFA) with groups defined by unordered categorical (i.e., nominal) between-subject variables. In MGCFA, MI is tested at each constraint of the latent factor model using a series of nested (latent) factor models. This traditional way of testing for MI originated with Jöreskog (1971), who was the first scholar to thoroughly discuss the invariance of latent factor (or measurement) structures. Additionally, Sörbom (1974, 1978) pioneered the specification and estimation of latent factor means using a multi-group SEM approach in LISREL (Jöreskog and Sörbom, 1996). Following these contributions the multi-group specification of latent factor structures has become widespread in all major SEM software programs (e.g., AMOS Arbuckle, 2006, EQS Bender and Wu, 1995, LAVAAN Rosseel, 2012, Mplus Muthén and Muthén, 2013, STATA STATA, 2015, and OpenMx Boker et al., 2011). Shortly thereafter, Byrne et al. (1989) introduced the distinction between full and partial MI. Although their introduction was of great value, the first formal treatment of different forms of MI and their consequences for the validity of multi-group/multi-time comparisons is attributable to Meredith (1993). So far, a tremendous amount of papers dealing

OPEN ACCESS

Edited by:
Jason W. Osborne, University of Louisville, USA

Reviewed by:
Guido Alessandri, Sapienza University of Rome, Italy

***Correspondence:**
Rens Van De Schoot, a.g.j.vandeschoot@uu.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *19 May 2015*
Accepted: *13 July 2015*
Published: *28 July 2015*

Citation:
Van De Schoot R, Schmidt P, De Beuckelaer A, Lek K and Zondervan-Zwijnenburg M (2015) Editorial: Measurement Invariance. Front. Psychol. 6:1064. doi: 10.3389/fpsyg.2015.01064

with MI have been published. The literature on MI published in the 20th century is nicely summarized by Vandenberg and Lance (2000). Noteworthy is also the overview of applications in cross-cultural studies provided by Davidov et al. (2014), as well as a recent book by Millsap (2011) containing a general systematic treatment of the topic of MI. The traditional MGCFA approach to MI-testing is described by, for example, Byrne (2004), Chen et al. (2005), Gregorich (2006), van de Schoot et al. (2012), Vandenberg (2002) and Wicherts and Dolan (2010). Researchers entering the field of MI are recommended to first consult Meredith (1993) and Millsap (2011) before reading other valuable academic works.

Recent developments in statistics have provided new analytical tools for assessing MI. The aim of this special issue is to provide a forum for a discussion of MI, covering some crucial "themes": (1) ways to assess and deal with measurement non-invariance; (2) Bayesian and IRT methods employing the concept of approximate MI; and (3) new or adjusted approaches for testing MI to fit increasingly complex statistical models and specific characteristics of survey data.

Dealing with Measurement Non-invariance

If the test for MI indicates that strict MI across groups or time is not established, no sound psychometric basis is provided for the comparison of latent factor means. The absence of such psychometric basic is the first topic dealing with measurement non-invariance. A nice example of a situation in which such psychometric basis is absent is provided in the paper by Lommen et al. (2014). These authors show that comparing posttraumatic stress in soldiers before and after war-zone related traumatic events (the wars in Afghanistan or Iraq) is virtually impossible due to instability in thresholds. For a researcher this conclusion may be hard to digest, especially if the success of the study relies entirely on the possibility to make such meaningful comparisons over time. Within the context of their study the authors recommend considering pre- and post-symptom scores as representing separate constructs.

In the same vein, a failure to establish less strict forms of MI may be worrisome if meaningful comparisons of structural relationships between latent factor means are important to the study (e.g., the comparison of the magnitude of a correlation, regression, or path coefficient across groups/time). Hox et al. (2015), show how the non-establishment of less strict forms of MI can (partly) be explained and corrected for. They show that, in the context of mixed-mode surveys, non-invariance can be the effect of selection or measurement differences due to mode (e.g., web survey, telephone survey, face-to-face interview).

Detecting non-invariant items is the next topic dealing with measurement non-invariance. In the contribution of de Roover et al. (2014) a method is proposed based on cluster-wise simultaneous component analysis (SCA). Their method aims at detecting non-invariant items. Barendse et al. (2014) examined a Bayesian restricted (latent) factor analysis (RFA) method for the same purpose, namely detecting items violating the MI assumption. They concluded that Bayesian RFA methods are especially suited for detecting measurement bias.

Our special issue also contains a discussion on the importance of understanding whether the presence of (in)correctly specified factorial invariance parameters influences the assessment of other factor model parameters (e.g., intercepts, error variances, latent factor variances, and latent factor means). In a simulation study, Guenole and Brown (2014) investigated whether ignoring the non-invariant underlying structure of the latent factor leads to substantial regression parameter bias in categorical item factor analyses (CIFA). The authors urge researchers to avoid ignoring sources of non-invariance in CIFA when non-invariance occurs in *both* loadings and thresholds even if this occurs in only one item.

Approximate Measurement Invariance

A relatively new research avenue in the MI literature deals with the use of Bayesian structural equation models (BSEM) to relax strict forms of MI (see Muthén and Asparouhov, 2012). In particular, exact zero constraints on the cross-group differences between all relevant measurement parameters (e.g., factor loadings and item intercepts) are substituted by "approximate" zero constraints. Instead of forcing item intercepts to be exactly equal across groups, a substantive prior distribution (around zero) is used to bring the parameters closer to one another, while allowing for some "wiggle room." If there are many small differences between the groups in terms of intercepts or factor loadings, approximate MI seeks a balance between adherence to the requirements of MI, making comparisons possible, and obtaining a well-fitting model (i.e., a model that is more realistic given the data at hand). When the classical MI tests do not hold given the data, approximate MI represents a promising (and more realistic) alternative; the cross-group differences between all relevant measurement parameters are "hopefully" close enough to zero to allow making meaningful latent factor mean comparisons.

A tutorial paper introducing the method of approximate MI is presented by van de Schoot et al. (2013). Further, our special issue contains empirical examples comparing the results of Bayesian approximate MI to the results of the more traditional ways of MI-testing as applied to specific questionnaires: e.g., the Portrait Values Questionnaire, using data from the European Social Survey including data on many countries and many time points (Cieciuch et al., 2014; Zercher et al., 2015), the Hedonic and Eudaimonic Motives for Activities scale (Bujacz et al., 2014), and the Golombok-Rust Inventory of Marital State (Chiorri et al., 2014).

Furthermore, our special issue contains two extensions of approximate MI to the field of IRT (see also Fox and Verhagen, 2010). Instead of using substantive prior distributions as in the Bayesian approximate MI method, the method described by Fox establishes a measurement scale across countries and conceptualizes country-specific non-invariance in item parameters as random deviations through country-specific random item effects. In such conceptualization cross-group comparisons can still be made even in the presence of non-invariant items. Kelcey et al. (2014) developed a method based on Fox's approximate MI approach which is applicable

whenever measurements are nested within raters and cross-classified among, for instance, countries. Another contribution to our special issue by Muthén and Asparouhov (2014) concerns the use of the *alignment* method (see also Asparouhov and Muthén, 2014) in IRT models, a method which is essential when applying approximate MI. This method minimizes a loss function which makes sure that there are a few large non-invariant measurement parameters instead of many smaller non-invariant measurement parameters, an optimal alignment strategy which resembles the rationale underlying rotation of factor solutions in EFA.

Testing for MI in Increasingly Complex Statistical Models

For some complex statistical models, the traditional multi-group (MGCFA) approach to MI-testing has to be adjusted to meet the specific requirements of the data and/or the model. Examples of such adjustments can be found in our special issue. An assumption embedded within many methods to test for MI is that the grouping (i.e., auxiliary) variable is unordered (i.e., nominal). Wang et al. (2014) present a method to test for MI in cases in which the auxiliary variable is ordered or continuous. Verdam and Oort (2014) illustrate MI-testing for Kronecker restricted SEM models, which constitute parsimonious models that provide an alternative to longitudinal latent factor models. Adolf et al. (2014) examine MI in the context of multiple-occasion and multiple-subject time series models. In such models, MI has to be established (a) over time within subjects, (b) over subjects within occasions, and (c) over time and subjects simultaneously. Boom (2014) investigated MI in the context of children's development of increasingly advanced strategies over time, in for instance the way they deal with mathematical problems (e.g., strategies on how children learn to multiply numbers below 10). The use of different strategies is scored as a variable and development is seen as the movement from one strategy to a more advanced one and Boom shows how MI plays a crucial role when analyzing such data. Jak (2014) uses a multi-level framework and proposes an extension to the SEM framework, moving from models describing two-level data to models describing three-level data. Within this framework MI invariance can be tested across level 2 as well as across level 3 clustering variables.

Another application of MI finds its origin in multi-trait multi-method models (MTMM; Eid and Diener, 2006), in which multiple methods (or scales) and raters are used to quantify the set of latent factors under study. Geiser et al. (2014) demonstrate the advantage of moving from an exclusively covariance- or correlation-based MTMM approach to an approach that includes latent factor means. This approach results in more fine-grained information about convergent validity and method effects when testing for MI. Albeit being analyzed differently, a comparable design to the MTMM is the two-way rating design utilized in situations where subjects have to judge to what extent a particular scale or variable pertains to a particular concept or situation. Kroonenberg (2014) presents an approach applicable to the assessment of MI in two-way rating designs. In his approach, a hierarchy of models is proposed, each one conceptualizing a form of MI, varying in terms of strictness.

Conclusion

Our special issue contains numerous simulation studies aiming at demonstrating the possibilities and limitations of different analytical tools to test for various forms of MI; tutorial papers providing the hands-on support needed when using the recent developed analytical tools to test for MI, as well as illustrations of how the analytical tools may be meaningfully applied in different fields of research when addressing issues related to MI across groups or time.

Acknowledgments

The first author was supported by a grant from The Netherlands Organization for Scientific Research: NWO-VENI-451-11-008. The second author was supported by the basic research program of the International Laboratory for Socio-Cultural Research at HSE Moscow.

References

Adolf, J., Schuurman, N., Borkenau, P., Borsboom, D., and Dolan, V. (2014). Measurement invariance within and between subjects: a distinct problem in testing the equivalence of intra- and inter-individual model structures. *Front. Psychol.* 5:883. doi: 10.3389/fpsyg.2014.00883

Arbuckle, J. L. (2006). *Amos (Version 7.0) [Computer Program]*. Chicago, IL: SPSS.

Asparouhov, T., and Muthén, B. (2014). Multiple-group factor analysis alignment. *Struct. Equ. Model.* 21, 1–14. doi: 10.1080/10705511.2014.919210

Barendse, M. T., Albers, C. J., Oort, F. J., and Timmerman, M. E. (2014). Measurement bias detection through Bayesian factor analysis. *Front. Psychol.* 5:1087. doi: 10.3389/fpsyg.2014.01087

Bender, P. M., and Wu, E. J. C. (1995). *EQS for Windows User's Guide*. Encino, CA: Multivariate Software.

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2011). OpenMx: an open source extended structural equation modeling framework. *Psychometrika* 76, 306–317. doi: 10.1007/s11336-010-9200-6

Boom, J. (2014). A new visualization and conceptualization of categorical longitudinal development: measurement invariance and change. *Front. Psychol.* 6:289. doi: 10.3389/fpsyg.2015.00289

Bujacz, A., Vittersø, J., Huta, V., and Kaozmarek, L. D. (2014). Measuring hedonia and eudaimonia as motives for activities: cross-national investigation through traditional and Bayesian structural equation modeling. *Front. Psychol.* 5:984. doi: 10.3389/fpsyg.2014.00984

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: a road less traveled. *Struct. Equ. Model.* 11, 272–300. doi: 10.1207/s15328007sem1102_8

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

Chen, F. F., Sousa, K. H., and West, S. G. (2005). Teacher's corner: testing measurement invariance of second-order factor models. *Struct. Equ. Model.* 12, 471–492. doi: 10.1207/s15328007sem1203_7

Chiorri, C., Day, T., and Malmberg, L.-E. (2014). An approximate measurement invariance approach to within-couple relationship quality. *Front. Psychol.* 5:983. doi: 10.3389/fpsyg.2014.00983

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., and Schwartz, S. H. (2014). Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Front. Psychol.* 5:982. doi: 10.3389/fpsyg.2014.00982

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., and Billiet, J. (2014). Measurement equivalence in cross-national research. *Annu. Rev. Sociol.* 40, 55–75. doi: 10.1146/annurev-soc-071913-043137

de Roover, K., Timmerman, M. E., DeLeersnyder, J., Mesquita, B., and Ceulemans, E. (2014). What's hampering measurement invariance: detecting non-invariant items using clusterwise simultaneous component analysis. *Front. Psychol.* 5:604. doi: 10.3389/fpsyg.2014.00604

Eid, M., and Diener, E. (2006). *Handbook of Multimethod Measurement in Psychology*. Washington, DC: American Psychological Association.

Fox, J.-P., and Verhagen, A. J. (2010). "Random item effects modeling for cross-national survey data," in *Cross-cultural Analysis: Methods and Applications*, eds E. Davidov, P. Schmidt, and J. Billiet (London: Routledge Academic), 467–488.

Geiser, C., Burns, G. L., and Servera, M. (2014). Testing for measurement invariance and latent mean differences across methods: interesting incremental information from multitrait-multimethod studies. *Front. Psychol.* 5:1216. doi: 10.3389/fpsyg.2014.01216

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Med. Care* 44(11 Suppl 3), 78–94. doi: 10.1097/01.mlr.0000245454.12228.8f

Guenole, N., and Brown, A. (2014). The consequence of ignoring measurement invariance for path coefficients in structural equation models. *Front. Psychol.* 5:980. doi: 10.3389/fpsyg.2014.00980

Hox, J. J., de Leeuw, E. D., and Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Front. Psychol.* 6:87. doi: 10.3389/fpsyg.2015.00087

Jak, S. (2014). Testing strong factorial invariance using three-level structural equation modeling. *Front. Psychol.* 5:745. doi: 10.3389/fpsyg.2014.00745

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika* 36, 109–133.

Jöreskog, K. G., and Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Uppsala: Scientific Software International.

Kelcey, B., McGinn, D., and Hill, H. (2014). Approximate measurement invariance in cross-classified rater-mediated assessments. *Front. Psychol.* 5:1469. doi: 10.3389/fpsyg.2014.01469

Kroonenberg, P. M. (2014). Assessing factorial invariance of two-way rating designs using three-way methods. *Front. Psychol.* 5:1495. doi: 10.3389/fpsyg.2014.01495

Lommen, M. J. J., van de Schoot, R., and Engelhard, I. M. (2014). The experience of traumatic events disrupts the stability of a posttraumatic stress scale. *Front. Psychol.* 5:1304. doi: 10.3389/fpsyg.2014.01304

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Muthén, B. O., and Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Front. Psychol.* 5:978. doi: 10.3389/fpsyg.2014.00978

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

Muthén, B. O., and Muthén, L. K. (2013). *Mplus Version 7.11 Statistical Analysis with Latent Variables: User's Guide*. Los Angeles, CA: Muthén & Muthén.

Rosseel, Y (2012). lavaan: an R package for structural equation modeling. *J. Stat. Software* 48, 1–36.

STATA. (2015). *Structural Equation Modeling Reference Manual 2015.* College Station, TX: Stata Press.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *Br. J. Math. Stat. Psychol.* 27, 229–239. doi: 10.1111/j.2044-8317.1974.tb00543.x

Sörbom, D. (1978). An alternative to the methodology for analysis of covariance. *Psychometrika* 43, 381–396. doi: 10.1007/BF02293647

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organ. Res. Methods* 5, 139–158. doi: 10.1177/1094428102005002001

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthen, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *Eur. J. Dev. Psychol.* 9, 486–492. doi: 10.1080/17405629.2012.686740

Verdam, M. G. E., and Oort, F. J. (2014). Measurement bias detection with Kronecker product restricted models for multivariate longitudinal data when the number of measurement occasions is large. *Front. Psychol.* 5:1022. doi: 10.3389/fpsyg.2014.01022

Wang, T., Merkle, E., and Zeileis, A. (2014). Score-based tests of measurement invariance: use in practice. *Front. Psychol.* 5:438. doi: 10.3389/fpsyg.2014.00438

Wicherts, J. M., and Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: an illustration using IQ test performance of minorities. *Educ. Measure.* 29, 39–47. doi: 10.1111/j.1745-3992.2010.00182.x

Zercher, F., Schmidt, P., Cieciuch, J., and Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: exact versus approximate measurement equivalence. *Front. Psychol.* 6:733. doi: 10.3389/fpsyg.2015.00733

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 September 2014
doi: 10.3389/fpsyg.2014.00883

Measurement invariance within and between individuals: a distinct problem in testing the equivalence of intra- and inter-individual model structures

Janne Adolf [1]***, Noémi K. Schuurman*** [2]***, Peter Borkenau*** [3]***, Denny Borsboom*** [4] ***and Conor V. Dolan*** [5]

[1] *Center for Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany*
[2] *Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Utrecht University, Utrecht, Netherlands*
[3] *Personality and Diagnostics Group, Department of Psychology, Faculty of Philosophy I, Martin-Luther-University Halle-Wittenberg, Halle, Germany*
[4] *Psychological Methods Group, Department of Psychology, Faculty of Social and Behavioral Sciences, University of Amsterdam, Amsterdam, Netherlands*
[5] *Department of Biological Psychology, Faculty of Psychology and Education, Free University of Amsterdam, Amsterdam, Netherlands*

Edited by:
Peter Schmidt, International Laboratory for Socio-Cultural Research HSE Moscow, Russia

Reviewed by:
Jan Boom, Utrecht University, Netherlands
Ellen Hamaker, Utrecht University, Netherlands
Henk Kelderman, Leiden University, Netherlands

****Correspondence:***
Janne Adolf, Center for Lifespan Psychology, Max Planck Institute for Human Development, Lentzeallee 94, Berlin 14195, Germany
e-mail: adolf@mpib-berlin.mpg.de

We address the question of equivalence between modeling results obtained on intra-individual and inter-individual levels of psychometric analysis. Our focus is on the concept of measurement invariance and the role it may play in this context. We discuss this in general against the background of the latent variable paradigm, complemented by an operational demonstration in terms of a linear state-space model, i.e., a time series model with latent variables. Implemented in a multiple-occasion and multiple-subject setting, the model simultaneously accounts for intra-individual and inter-individual differences. We consider the conditions—in terms of invariance constraints—under which modeling results are generalizable (a) over time within subjects, (b) over subjects within occasions, and (c) over time and subjects simultaneously thus implying an equivalence-relationship between both dimensions. Since we distinguish the measurement model from the structural model governing relations between the latent variables of interest, we decompose the invariance constraints into those that involve structural parameters and those that involve measurement parameters and relate to measurement invariance. Within the resulting taxonomy of models, we show that, under the condition of measurement invariance over time and subjects, there exists a form of structural equivalence between levels of analysis that is distinct from full structural equivalence, i.e., ergodicity. We demonstrate how measurement invariance between and within subjects can be tested in the context of high-frequency repeated measures in personality research. Finally, we relate problems of measurement variance to problems of non-ergodicity as currently discussed and approached in the literature.

Keywords: measurement invariance, ergodicity, state-space modeling, latent variables, intra-individual level of analysis

INTRODUCTION

Population heterogeneity exists when multiple distinct statistical models are required to adequately describe a population (Muthén, 1989). Statistical approaches to investigate and accommodate heterogeneity include, for instance, multi-group modeling (e.g., Jöreskog, 1971; Muthén, 1989), multi-level modeling (e.g., Hox, 2002), and structural equation mixture modeling (e.g., Dolan, 2009). In each of these modeling approaches a heterogeneous population is stratified into subpopulations whose members adhere to the same models and differences within are separated from differences between subpopulations (Muthén, 1989). But how small is the smallest subgroup? One could think of a scenario in which breaking up a heterogeneous population into ever smaller subpopulations leads to the smallest subpopulation that is empirically realizable. This is the individual person (Millsap, 2011). Consider, for instance, the five-factor-model (FFM) which states that the dimensions Extraversion, Neuroticism, Agreeableness, Conscientiousness and Openness to Experience are the major sources of inter-individual differences in personality (McCrae and John, 1992). A researcher studying population heterogeneity can now well question, whether the FFM is generally interpretable in the sense that it holds for each individual member of the overall population by addressing "universal" determinants of human behavior (Hamaker et al., 2005).

Questions of this kind have indeed been posed recently and have been addressed by means of single subject ($N = 1$) modeling based on the analysis of repeated measurements over occasions (Cattell, 1952; Gregson, 1983; Molenaar, 1985). By contrasting intra-individual with inter-individual difference data, it has been shown that inter-individual modeling results do usually not generalize to the level of the individual. Rather, individual specifics, which remain undetected in standard large sample modeling techniques, seem to be the rule, not the exception (e.g., Molenaar et al., 2003; Molenaar, 2004; Hamaker

et al., 2005, 2007; Kelderman and Molenaar, 2007; Molenaar and Campbell, 2009; Schmiedek et al., 2009; Brose et al., 2010, 2014; Nesselroade, 2010). The increasing interest in individual modeling techniques therefore emphasizes the conceptual continuity between approaches to heterogeneous populations and to the individual. Explicitly stated, single subject modeling accommodates population heterogeneity in its most extreme sense as it does not necessarily involve the generalization of results to other individuals or subpopulations of individuals. Each individual can thus potentially represent a system that is quantitatively or qualitatively unique (Molenaar, 2004).

We have so far conceived of heterogeneity as heterogeneity between individuals, but one may just as well conceive of heterogeneity as heterogeneity within individuals. That is, an individual's system characteristics may display (higher order) stability or variability over time (Molenaar, 2004). To illustrate this, suppose a researcher aims at describing a person with respect to a certain attribute over time. One may now think of an intra-individual distribution of states rather than of a single trait score. Considered over a representative set of situations, this distribution may have relatively stable characteristics over time, e.g., stable mean and variance. These may then be used to differentiate among people and may thus themselves be regarded as personality characteristics (Fleeson, 2001; Hamaker et al., 2007). However, also within individuals, homogeneity cannot be taken for granted but constitutes a (restrictedly) testable assumption. Similarly to questioning to what extent population models generalize to individual population members, one could question to what extent an individual time series model generalizes to (subsets of) single occasions.

The reorientation toward the individual in differential psychology has been motivated by and motivates an integrative consideration of the within- and the between-subject perspective. It therefore provides an optimal setting to address the following guiding questions: Under what conditions are modeling results generalizable (a) over occasions within subjects, (b) over subjects within occasions, and (c) over occasions and subjects simultaneously? Question (c) refers to the conditions that establish a systematic relationship, i.e., equivalence between the structure of intra- and the structure of inter-individual data (given large N and T). Borrowing terminology from statistical mechanics, this situation is termed *ergodicity* in the psychometric literature (e.g., Molenaar et al., 2003; Molenaar, 2004; Molenaar and Campbell, 2009). In the present context, ergodicity is referred to as a situation in which the statistical behavior of a time series observed for a single subject is the same as the statistical behavior of a sample of multiple subjects, obtained at a few occasions (i.e., the definition of an ergodic process according to Molenaar, 2004, p. 208).

Psychological attributes, however, are often represented as latent variables, the study of which requires psychometric measurement. In the context of latent variable modeling the conditions for an ergodic process decompose into invariance constraints on the structural part of the model and invariance constraints on the measurement model. The latter constraints relate to the concept of measurement invariance (MI; Mellenbergh, 1989; Meredith, 1993; Millsap, 2011). In this paper, we discuss how MI ties into the integrated within- and between-subject context. Specifically, we focus on how the concept is to be considered when one is interested in investigating the generalizability of latent variable modeling results along the dimensions time and subject.

The outline of the paper is as follows. Based on the definition as provided by Mellenbergh (1989), we elaborate on MI in the between- and within-subject context, in general terms and operationally in the linear factor model which lends itself well to integrated modeling, i.e., simultaneous modeling of intra- and inter-individual differences. We then proceed to address our guiding questions using a bottom-up approach. That is, in a multiple-subject, multiple-occasion setting, we set up a linear multi-subject latent variable time series model that accounts for intra-individual and inter-individual variability and we implement the model constraints that imply generalizability of results along the dimensions time and subject. We consider these constraints separately at the level of the measurement process and at the level of the latent psychological process. The result is a taxonomy of differently restrictive models ranging from full heterogeneity to full homogeneity between and within individuals. It can be considered a taxonomy of problems[1] a researcher will potentially face when simultaneously modeling intra- and inter-individual variation. We show that MI holding simultaneously over time and subject can be interpreted as constituting a mode of structural equivalence between the intra- and the inter-individual level of analysis that is distinct from full structural equivalence. Using a real data illustration on intra-individual variability in the personality domain (Borkenau and Ostendorf, 1998), we show how researchers can test for MI over subjects and time. In the discussion, we reconsider the assumptions underlying MI testing and review alternative interpretations of and potential approaches to measurement variance within and between subjects.

MEASUREMENT INVARIANCE BETWEEN AND WITHIN SUBJECTS

GENERAL DEFINITION OF MEASUREMENT INVARIANCE

The present focus on MI is motivated by the latent variable paradigm which informs conceptual thinking in modern psychology (Bollen, 2002; Borsboom et al., 2003; Borsboom, 2008; Millsap, 2011). Although not directly observable, an attribute such as agreeableness can be conceptualized as manifesting in terms of observable behaviors or reportable attitudes, in this case along the interpersonal dimensions warmth, kindness, appreciation, and consideration (McCrae and John, 1992; Graziano and Tobin, 2009). However, inferences about latent variables on basis of observed indicators are subject to relatively large uncertainty (Borsboom, 2008). MI is one of the psychometric concepts addressing this uncertainty.

A general formal definition of MI in the latent variable paradigm was given by Mellenbergh (1989). Suppose we have a set of indicators $\boldsymbol{Y}$ that together form a psychometric instrument

[1] This useful notion was suggested by one of the reviewers.

designed to measure a given latent variable Z, and suppose we have a variable X. MI of the indicators with respect to X is defined as independence of the indicators and X conditional on the latent variable, i.e.,

$$f(\boldsymbol{Y}|Z = z) = f(\boldsymbol{Y}|Z = z,\ X = x) \quad (1)$$

for all values of Z and X, in which $f(\cdot)$ denotes the probability distribution function. Under MI, any effect of X on the indicators is indirect, i.e., mediated through the latent variable (Lubke et al., 2003b). Consequently, significant differences in observed indicator scores are attributable to differences in the targeted latent variable (Z) across units selected on basis of X, e.g., across persons (e.g., Mellenbergh, 1989; Horn and McArdle, 1992; Lubke et al., 2003b; van der Sluis et al., 2006; Wicherts and Dolan, 2010; Millsap, 2011).

To illustrate this, imagine we attempted to measure agreeableness (Z) in a given sample using questionnaire $\boldsymbol{Y}$. Let X be the tendency to respond in a socially desirable manner (Paulhus and Reid, 1991; Holtgraves, 2004). If $\boldsymbol{Y}$ was measurement invariant with respect X, any two individuals from the sample having the same level of agreeableness would attain the same score on each item (apart from measurement error effects). Importantly, they would do so independent of their potentially different tendencies to respond in a socially desirable manner. $\boldsymbol{Y}$ would then be considered unbiased with respect to X. On the contrary, if $\boldsymbol{Y}$ was measurement variant or biased with respect to X, for instance due to item contents triggering socially desirable responding, differences in individual's responses would not necessarily be interpretable as differences in agreeableness. They may as well be interpretable as differences in socially desirable responding. Measurement variance or bias thus refers to a replicable difference in item scores which is not due to the targeted latent variable Z (Millsap, 2011). Meaningful comparisons in terms of the targeted latent variable are thus not guaranteed on basis of biased item scores (e.g., Dolan et al., 2004; Hamaker, 2007; Raykov et al., 2012).

Moreover, biased items can lead to biased estimates of parameters pertaining to the latent variable (Mellenbergh, 1989; Wicherts and Dolan, 2010). The interpretation of the latent variable is then rendered problematic. The converse argument would be that, if MI across persons selected on basis of X holds, the interpretation of the latent variable is the same across these persons (e.g., Mellenbergh, 1989; Horn and McArdle, 1992; Lubke et al., 2003a; Dolan et al., 2004; Borsboom and Dolan, 2007; Nesselroade et al., 2007; Wicherts and Dolan, 2010; Raykov et al., 2012). This notion of MI as *theoretical invariance*, as compared to the above notion of *unbiasedness*, can mainly be found for operationalizations of MI in the linear factor model. It is argued that the interpretation of the factor is determined by its relation to the observed indicators (the factor loadings) and that it is unlikely that different factors are related to a fixed set of indicators in exactly the same way (Lubke et al., 2003a).

Regardless of which interpretational notion is employed, in applying the concept of MI, one has to rely on premises which may appear more or less sensible depending on the context. We get back to this in more detail in the discussion.

CONCEPTUALIZATION OF MEASUREMENT INVARIANCE BETWEEN AND WITHIN SUBJECTS

MI has been investigated extensively in the context of multi-group factor analysis, with groups defined by nominal between-subject variables, such as sex or ethnic background (e.g., van der Sluis et al., 2006; Wicherts and Dolan, 2010). Mellenbergh's definition, however, is a general one. It is neutral with respect to the nature and format of the potentially biasing variable, the indicator variables, and latent variables, and is thus independent of the psychometric model that relates the indicators to the latent variables (Mellenbergh, 1989; Meredith, 1993; Lubke et al., 2003a; Wicherts and Dolan, 2010). We can therefore draw two conclusions in the present context. First, Mellenbergh's definition should be equally applicable at the between-subject and at the within-subject level (Borsboom and Dolan, 2007). MI can also be considered with respect to time-varying variables relevant within subjects, such as mood or work pressure. For instance, a questionnaire supposed to assess intra-individual fluctuations in the state agreeableness over time may be biased with respect to mood. Then, a person's series of responses over time would reflect not only variations in the state agreeableness but additionally variations in mood. The second conclusion based on Mellenbergh's general definition is, that it is possible to take a more general perspective and consider MI with respect to subject and time (index) itself. This relates back to our introductory questions[2].

OPERATIONALIZATION OF MEASUREMENT INVARIANCE BETWEEN AND WITHIN SUBJECTS

Mellenbergh's general MI definition gives rise to testable model constraints when implemented in the context of a concrete latent variable model. The latent variable modeling framework explicitly distinguishes between a (reflective) measurement model, in which the observed indicators are modeled as a function of the latent variables of psychological interest, and a structural model, which concerns the latent variables and their interrelationships. The linear factor model may be viewed as a proper measurement model in which multiple continuous indicators are linearly regressed upon a single continuous latent variable (e.g., Mellenbergh, 1994). In the linear factor model, MI has been associated with the constraints of strict factorial invariance (strict FI; Meredith, 1993) for the standard between-subject context. However, this measurement model features not only

[2]The shift in perspective from MI with respect to specific variables to MI over subjects or time has interesting implications (cf. Meredith, 1993, p. 529, theorem 3). MI over subjects implies MI with respect to any variable that varies exhaustively over subjects within the population considered. Equivalently, and under the assumption of an appropriate sampling rate over time, MI over time implies MI with respect to any variable that varies exhaustively within the period of time considered. Hence, by taking this perspective, one automatically accounts for all measured or unmeasured (discrete and finite) background variables that vary along the dimensions time and subject (cf. Lubke et al., 2003b).

in structural equation modeling at the between-subject level (SEM) but also in state-space modeling of time series data at the within-subject level (SSM; Oud et al., 1990; Chow et al., 2010). We argue that strict FI should be equally applicable at the inter-individual and the intra-individual level. That is, strict FI over (subsets of) subjects within occasions, i.e., subject invariant measurement parameters such as factor loadings, intercepts and residual variances should almost certainly imply MI over subjects within occasions. In addition, strict FI over (subsets of) occasions or time within subjects, i.e., time-invariant measurement parameters, should almost certainly imply MI over time within subjects for the given sampling rate[3].

A BOTTOM-UP APPROACH FROM FULL HETEROGENEITY TO ERGODICITY

THE BASELINE MODEL

We now demonstrate the relation between ergodicity and MI in the context of linear stochastic time series models in state-space format (Harvey, 1989; Oud et al., 1990; Hamilton, 1994; Durbin and Koopman, 2001; Hamaker and Dolan, 2009; Chow et al., 2010). Such models primarily account for intra-individual variation over time. However, by specifying them within many subjects simultaneously we can extend them to multi-subject models. The conditions under which modeling results are generalizable over time, over subjects, and over time and subjects simultaneously may then be expressed in terms of specific invariance constraints. Furthermore, the state-space format incorporates a measurement model and a latent process model which allows distinguishing among constraints that apply to the measurement parameters and constraints that apply to latent parameters. In the following, subscript i and t refer to subject and discrete time, respectively. We assume equidistant measurement occasions throughout.

The latent process model is formulated as

$$\boldsymbol{\eta}_{i,\,t} = \boldsymbol{\alpha}_{i,\,t} + \mathbf{B}_{i,\,t}\boldsymbol{\eta}_{i,\,t-1} + \boldsymbol{\zeta}_{i,\,t} \tag{2}$$

where $\boldsymbol{\eta}_{i,\,t}$ is a $q \times 1$ vector of latent variables, the states, which are regressed on themselves at the previous time point, $\mathbf{B}_{i,\,t}$ is a $q \times q$ matrix of latent regression parameters capturing the auto- and cross-lagged regression relationships among the states over time, and $\boldsymbol{\alpha}_{i,\,t}$ is a $q \times 1$ vector of latent regression intercepts. The vector $\boldsymbol{\zeta}_{i,\,t}$ is a $q \times 1$ vector of latent residuals which are assumed to be multivariate normally distributed with mean zero and covariance matrix $\boldsymbol{\Psi}_{i,\,t}$. The latent residuals are uncorrelated over time and uncorrelated with $\boldsymbol{\eta}_{i,\,t-1}$. The model-implied mean vector of the latent states, $\boldsymbol{\nu}_{i,\,t}$, can be expressed as a function of $\boldsymbol{\alpha}_{i,\,t}$, $\mathbf{B}_{i,\,t}$, and $\boldsymbol{\nu}_{i,\,t-1}$. The model-implied covariance-matrix of the latent states, $\mathbf{P}_{i,\,t}$, can be expressed as a function of $\mathbf{B}_{i,\,t}$, and $\mathbf{P}_{i,\,t-1}$ and $\boldsymbol{\Psi}_{i,\,t}$. Note that although the formal process is driven by a vector autoregressive process of first order, the actual psychological process needs not obey this structure. This so-called single lag structure renders the model fitting process technically convenient. However, any uni- or multivariate autoregressive moving average model can be accommodated (i.e., reformulated in terms of a first order vector autoregressive process) by extending the state vector by the relevant process components (e.g., Harvey, 1989; Hamaker and Dolan, 2009; Shumway and Stoffer, 2011).

The measurement model is formulated as

$$\mathbf{y}_{i,\,t} = \boldsymbol{\tau}_{i,\,t} + \boldsymbol{\Lambda}_{i,\,t}\boldsymbol{\eta}_{i,\,t} + \boldsymbol{\varepsilon}_{i,\,t} \tag{3}$$

where $\mathbf{y}_{i,\,t}$ is a $p \times 1$ vector of manifest indicators, $\boldsymbol{\Lambda}_{i,\,t}$ is a $p \times q$ matrix of factor loadings and $\boldsymbol{\tau}_{i,\,t}$ is a $p \times 1$ vector of measurement intercepts. The $p \times 1$ vector $\boldsymbol{\varepsilon}_{i,\,t}$ contains measurement residuals, ideally measurement errors, which are assumed to be multivariate normally distributed with mean zero and covariance matrix $\boldsymbol{\Theta}_{i,\,t}$. The measurement residuals are uncorrelated over time and uncorrelated with $\boldsymbol{\eta}_{i,\,t}$ and $\boldsymbol{\zeta}_{i,\,t}$. Here, we additionally assume zero correlations among the measurement residuals, i.e., $\boldsymbol{\Theta}_{i,\,t}$ is diagonal, satisfying the assumption of local independence. The model-implied mean vector of the indicators, $\boldsymbol{\mu}_{i,\,t}$ can be expressed as a function of $\boldsymbol{\tau}_{i,\,t}$, $\boldsymbol{\Lambda}_{i,\,t}$, and $\boldsymbol{\nu}_{i,\,t}$. The model-implied covariance-matrix of the indicators, $\boldsymbol{\Sigma}_{i,\,t}$, can be expressed as a function of $\boldsymbol{\Lambda}_{i,\,t}$, and $\mathbf{P}_{i,\,t}$ and $\boldsymbol{\Theta}_{i,\,t}$. As noted, this measurement model is equivalent to the linear factor model as it features in standard between-subject SEM (Oud et al., 1990; Chow et al., 2010).

The model in Equations (2) and (3) is our baseline model. Note that the model is completely unrestricted with respect to time and subject, meaning that all model parameters can vary in value over time and subjects, but also that the model structure can be subject- and time-dependent. This concerns the dimensionality of the state vector, the pattern of factor loadings, and in the pattern of interrelationships among latent states and latent residuals. As a consequence, the model-implied covariance matrix, and the model-implied mean vector are subject- and time-dependent. Theoretically, the model does thus accommodate full heterogeneity within and between subjects. We now impose increasingly restrictive invariance constraints relating to the dimensions time and subject. We first consider the model constraints that lead from total heterogeneity to MI over time and subjects. We then consider the additional model constraints that eventually result in full invariance over time and subjects, i.e., an ergodic process, as discussed by Molenaar and colleagues (e.g., Molenaar, 2004; Molenaar and Campbell, 2009).The different models are organized in form of a taxonomy. **Figure 1** represents this taxonomy in terms of model equations and verbal terms. As we are interested in the conditions that establish equivalence between the intra- and inter-individual level of analysis, we focus on those models in which we impose constraints simultaneously within and between subjects.

[3]Under the assumptions that multivariate normality holds, it is unlikely that variation in measurement error variance and variation in specific factor variance cancel each other out across occasions and subjects respectively, and it is unlikely that variation in measurement intercepts and variation in specific factor means cancel each other out across occasions and subjects respectively (cf. Meredith, 1993; Lubke et al., 2003a,b).

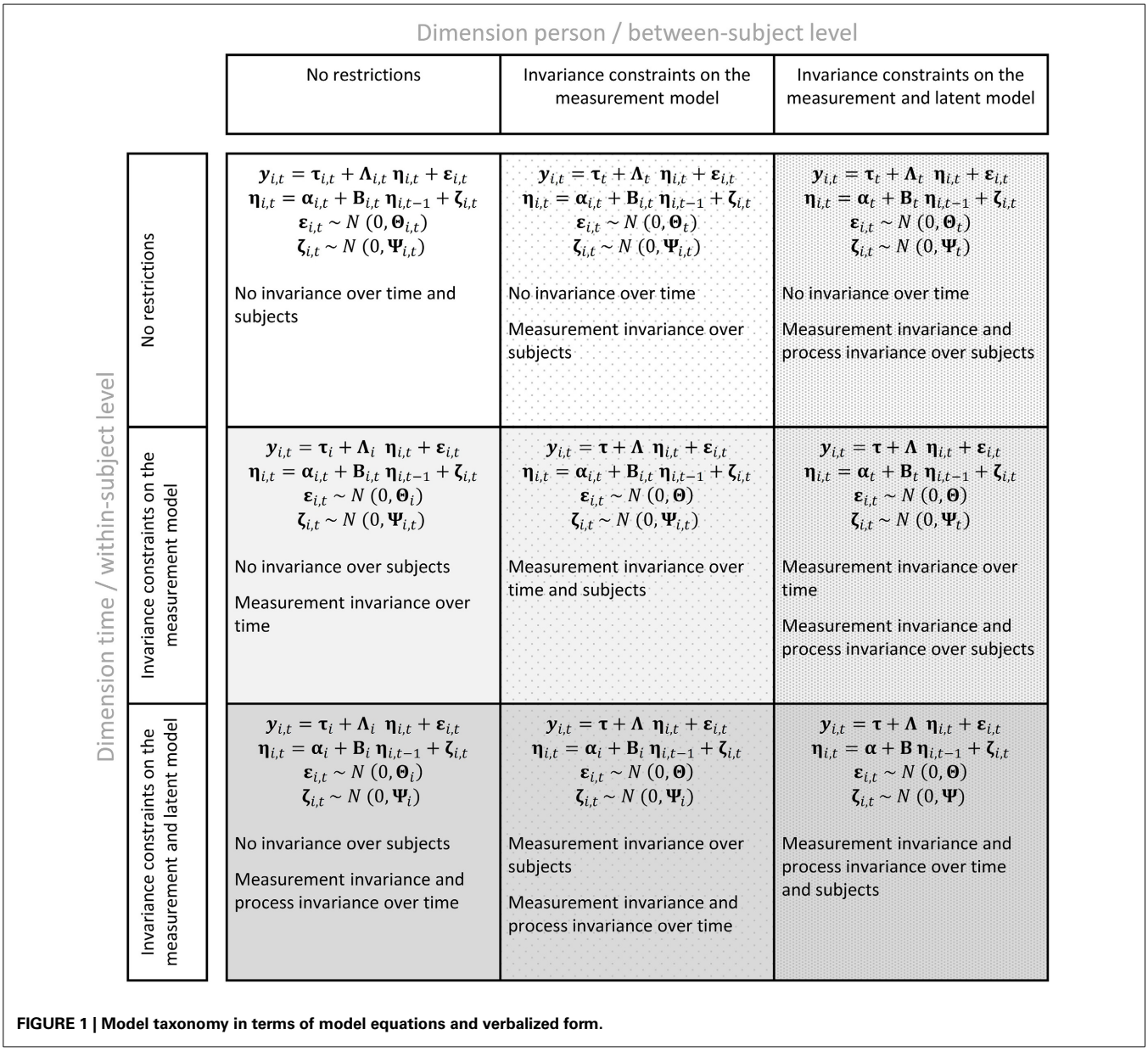


FIGURE 1 | Model taxonomy in terms of model equations and verbalized form.

MODES OF EQUIVALENCE BETWEEN THE INTRA- AND INTER-INDIVIDUAL LEVEL OF ANALYSIS

We first consider the baseline model as a reference. As presented in Equations (2) and (3) neither the measurement model nor the latent process model is restricted over time or over subjects. Note that, technically, the model is not identified until some sort of time-related pattern is imposed. Assuming some pattern would also be indicated from a theoretical perspective. This needs however not involve constraining (measurement) model parameters to be time-invariant. There is thus no equivalence relationship between the intra- and the inter-individual level. A model based on pooled data over occasions and subjects would address a process that is a mixture over time and subjects unconditional and conditional on the latent process (cf. Muthén, 1989). Applying the interpretation of MI as unbiasedness results in the following conclusions. The absence of MI over time within subjects due to time-varying measurement parameters indicates that within any given person there is systematic observed variability over time that is not attributable to the targeted latent variables in $\eta_{i,\,t}$. Since MI over subjects within time points does also not hold due to person-specific measurement parameters there is systematic observed variability between persons that is not attributable to the targeted latent variables. Different time- and subject-varying variables may cause measurement variance and these associations may be person- and indicator-specific and may change over time. As long as these (unknown) variables and their effects on the indicators are not accounted for, the interpretation of the latent variables as

they develop over time and differ over subjects remains complicated. This is in accordance with the notion of MI as theoretical equivalence which holds that the latent variables in $\eta_{i,\,t}$ are not necessarily interpretable in an invariant sense over time or subjects. That would become directly apparent in an extreme case, in which the measurement model would display different factor loading patterns over time or subjects. In the discussion, we elaborate on recently suggested strategies to handle and explore such a situation.

By constraining all parameters to be invariant over time and subjects we obtain the extreme opposite. The measurement and process model reduce to

$$\mathbf{y}_{i,\,t} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta}_{i,\,t} + \boldsymbol{\varepsilon}_{i,\,t} \tag{4}$$

and

$$\boldsymbol{\eta}_{i,\,t} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_{i,\,t-1} + \boldsymbol{\zeta}_{i,\,t} \tag{5}$$

with

$$\begin{aligned} \boldsymbol{\varepsilon}_{i,\,t} &\sim N(0,\ \boldsymbol{\Theta}), \\ \boldsymbol{\zeta}_{i,\,t} &\sim N(0,\ \boldsymbol{\Psi}). \end{aligned}$$

An additional requirement ensuring stationarity of the latent process, i.e. time-invariant process characteristics, is that all eigenvalues of matrix **B** are less than one in absolute value (Hamilton, 1994; Molenaar, 2004). Note that the model-implied distributions of observed and latent variables are now independent of subject and time. This model thus represents an operationalization an ergodic process under the assumption of normality (Molenaar, 2004, p. 208). Under these conditions one (intra-individual) process model generalizes across the entire time span and across all subjects in the population considered, i.e., the individual state-space time series models coincide with a standard between-subject longitudinal factor model based on at least two occasions (Molenaar et al., 2003; Molenaar, 2004). Consequently, the between-subject model provides a description of the intra-individual dynamics of each individual in the population and over the entire period of time considered (e.g., Molenaar, 2004; Hamaker et al., 2005; Molenaar and Campbell, 2009). Pooling over persons and time points is feasible as modeling results are fully generalizable between and within subjects.

Between these two extreme variants is the model in which the invariance constraints only concern the measurement model. Strict FI imposed simultaneously with respect to time and subject implies MI with respect to time and subject and results in the model

$$\mathbf{y}_{i,\,t} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta}_{i,\,t} + \boldsymbol{\varepsilon}_{i,\,t} \tag{6}$$

and

$$\boldsymbol{\eta}_{i,\,t} = \boldsymbol{\alpha}_{i,\,t} + \mathbf{B}_{i,\,t}\boldsymbol{\eta}_{i,\,t-1} + \boldsymbol{\zeta}_{i,\,t} \tag{7}$$

with

$$\begin{aligned} \boldsymbol{\varepsilon}_{i,\,t} &\sim N(0, \boldsymbol{\Theta}), \\ \boldsymbol{\zeta}_{i,\,t} &\sim N\left(0,\ \boldsymbol{\Psi}_{i,\,t}\right). \end{aligned}$$

Note that the conditions for MI over time and subjects concern only the measurement process, that is, invariance of the model parameters over time and subjects conditional on the latent process. Simultaneous MI over time and subjects thus represents a form of structural equivalence between levels of analysis that still allows for substantial heterogeneity with respect to the latent variables and their interrelations over time and over subjects. Consequently, we propose to distinguish between two *modes* of structural equivalence. That is, a mode of measurement equivalence, which involves MI over time and subjects but does not include equivalence of the interrelations among the latent variables and latent residuals, and a distinct mode of full equivalence, which is ergodicity. A model based on data pooled over occasions or subjects would imply a latent process that is a mixture over time and subjects whereas modeling results regarding the measurement process would be generalizable over time and subjects.

Interpreting MI as biasedness of the indicators, this model implies that systematic observed intra-individual as well as inter-individual variability is attributable to the targeted latent variables in $\boldsymbol{\eta}_{i,\,t}$. The interpretation as theoretical invariance holds that the same latent variables are measured within and between subjects. Systematic within- and between-subject variation can be viewed as variation on the same set of latent variables (cf. Lubke et al., 2003a). The model would thus capture intra-individual dynamics and inter-individual differences therein with respect to the targeted latent variables (cf. Hamaker et al., 2007). In this sense, measurement equivalence could be considered a necessary condition for studying intra- and inter-individual differences pertaining to the latent variables of interest.

ILLUSTRATION

PURPOSE OF ILLUSTRATION, DATA DESCRIPTION, AND SELECTION

We show how measurement invariance can be investigated (a) over subjects and (b) over time within a given subject. As we use a modeling approach for stationary time series data we shall limit our illustration to time series models which we assume to be invariant with respect to time. We demonstrate below, that these models allow us to incorporate measurement variance over time to a limited extent.

We use data from Borkenau and Ostendorf (1998) that consist of individual time series of self-ratings on personality items. On 90 successive days, 22 students indicated the degree to which 30 adjectives applied to their daily state. Standard between-subject factor analysis showed that the items measure the inter-individual difference traits Neuroticism, Extraversion, Agreeableness, Conscientiousness and Openness to Experience (e.g., Borkenau and Ostendorf, 1990; McCrae and John, 1992; Borkenau and Ostendorf, 1998). The response format was a 7 point scale with high scores

indicating high correspondence between described and perceived state.

For our present illustration, we consider a subset of items and subjects with approximately continuously and normally distributed responses, and the absence of obvious mean-level-trends or variability-changes in the series over time[4]. We focus on three individuals (subjects 7, 13, and 22), and their responses to the extraversion ("dynamic," "sociable," "shy," "silent," "lively," "reserved") and agreeableness marker items ("selfish," "good-natured," "domineering," "helpful," "obstinate," "considerate"). The individual data and descriptive figures are available as supplementary materials.

DETERMINING THE INDIVIDUAL STATE-SPACE TIME SERIES MODELS

To set up the individual models, we imposed a two-factor measurement model on each individual's data, such that the extraversion marker items load on one, the agreeableness marker items on a second factor. Note that there is no guarantee that the two-factor model, which would be expected to fit the data in standard inter-individual factor analysis, will fit the individual time series data (e.g., Molenaar, 2004; Hamaker et al., 2005; Molenaar and Campbell, 2009). By means of exploratory factor analysis, one could identify individual factor solutions that would potentially be person-specific (regarding sets of factors and factor loading patterns) and then conduct within-person fit comparisons between the individual models and the two-factor model (e.g., Hamaker et al., 2005, 2007). Here, we assume configural invariance over individuals, that is, an invariant number of factors and an invariant factor loading pattern (Meredith, 1993).

We determined the individual process models by modeling the auto- and cross-lagged relationships among the factors using the Fortran program MKF (Dolan, 2010)[5]. This program can fit linear stochastic time series models in state-space format to stationary time series data via the linear, time-invariant Kalman filter algorithm. For correctly specified state-space models the Kalman filter provides optimal estimates of the latent variable states over time and gives rise to ML estimates of the model parameters. Detailed explanations of the estimation procedure can for instance be found in the econometric (e.g., Harvey, 1989; Hamilton, 1994; Durbin and Koopman, 2001) and psychometric literature (e.g., Oud et al., 1990; Chow et al., 2010). Within each individual we contrasted vector auto-regressive processes of first order (VAR(1)), second order (VAR(2)), and of order zero (VAR(0)). In the last case, the factors do not display lagged relationships. We pruned models by fixing to zero non-significant relationships in $\mathbf{B}_i$ and $\mathbf{\Psi}_i$ (overall-$\alpha = 0.05$). We imposed scaling by fixing the latent intercepts to zero and the latent residual variances to one. The information criteria BIC (Schwarz, 1978) and AIC (Akaike, 1974) served as main indicators for relative model fit but we also conducted Log-Likelihood difference tests where models were nested ($\alpha = 0.05$). **Table 1** provides an overview of the results and **Figure 2** shows path diagrammatic representations of the individual models.

According to AIC and BIC, subjects 7 and 22 both display a latent process that involves lagged relationships among the factors. For subject 7 there is only one auto-regressive effect of first order for the agreeableness factor, for subject 22 there is the full set of first- and second-order auto- and cross-lagged regression effects. In case of subject 13 the latent process does not contain any lagged effects among the factors. Within occasions, both factors are correlated within each of the three subjects.

With respect to the individual measurement models, the loadings relating the extraversion indicators to the corresponding factor seem to be relatively homogeneous and reasonably large within each individual (although the measurement residual variances are consistently large). This is different for the agreeableness indicators which are associated not only with more heterogeneous loadings but also with loadings close to zero as in case of the item "helpful." Especially for subject 7 it is questionable whether one coherent dimension underlies his or her responses to the agreeableness indicators. However, to test this we would have to employ a more explorative approach as outlined above. Note that the loading signs suggest that the factors are inverted in some cases.

ADDRESSING MI OVER SUBJECTS

To address MI over subjects we made use of the multi-group modus in MKF treating each individual as a group. FI was then tested via pairwise comparisons between all three subjects. Since we scaled in the latent space by standardizing the conditional latent states, all factor loadings and measurement intercepts are freely estimated and can thus all be subjected to a test of invariance across groups (Raykov et al., 2012). In order to not confound FI constraints with invariance constraints pertaining to the latent level, we freely estimated the latent residual variances in one of the subjects whenever the factor loadings were constrained to equality. Equivalently, we freed the latent intercepts in one of the models, whenever the measurement intercepts were constrained to equality (Wicherts and Dolan, 2010; Raykov et al., 2012). **Table 2** provides an overview of the results.

For all pairwise comparisons between subjects, the AIC and the BIC favored the weakly factorial invariant model. Note that a χ^2-difference-test for instance between the configurally invariant and the strictly factorial invariant model cannot be conducted as the models are not nested. This is due to the freely estimated latent parameters in the strictly factorial invariant model (Raykov et al., 2012). The finding of subject-invariant factor loadings suggests that the same dimensions underlie the variation within each of the three individuals (Hamaker et al., 2007). These are however not necessarily the dimensions underlying the differences between

[4]We selected subjects based on visual inspection of the frequency distributions and time series plots of their responses. Although the five factor marker items may be considered discrete, they are often treated as continuous in the literature (e.g., Borkenau and Ostendorf, 1998; Hamaker et al., 2005, 2007; Rammstedt and John, 2005). Indeed, Dolan (1994) demonstrated, that treating indicators with at least seven ordered response categories as continuous, does not affect standard errors and overall test statistics of normal theory maximum likelihood estimation—if the distribution of each indicator is not too skewed. Lubke and Muthén (2004) investigated problematic effects of skewed indicator distributions of pseudo-continuous items in standard confirmatory factor analysis.

[5]The program (including documentation) is available by request from c.v.dolan@vu.nl. All MKF in- and output files for the models fitted are available as supplementary materials. These also include R-code to set up data and input files for MKF, execute MKF, and read MKF output files.

Table 1 | Comparison of different process models within individuals.

Process model	npars	−2LogL	AIC	BIC	χ^2-increase (relative to)	*df*	*p*
Subject 7							
VAR (0)	37	1089	1163	1255	10.377 (VAR (1))	4	0.035
					6.993 (VAR (1)*)	1	0.008
VAR (1)	41	1079	1161	1263			
*VAR (1)**	38	1082	1158	1253	3.384 (VAR (1))	3	0.336
VAR (2)	45	1095	1185	1297			
Subject 13							
VAR (0)	37	1522	1596	1689	5.221 (VAR (1))	4	0.265
VAR (1)	41	1517	1599	1702			
VAR (2)	45	1515	1605	1718			
Subject 22							
VAR (0)	37	1212	1286	1378	23.655 (VAR (1))	4	0.000
VAR (0)*	36	1214	1286	1376	1.815 (VAR (0))	1	0.178
VAR (1)	41	1188	1270	1373			
VAR (1)*	37	1202	1276	1368	13.366 (VAR (1))	4	0.010
VAR (2)	45	1161	1251	1363.7			
VAR (2)*	39	1189	1267	1364.1	27.390 (VAR (2))	6	0.000

Model variants denoted with an asterisk are pruned with respect to simultaneous and lagged relationships. The relatively best fitting model according to AIC and BIC is set in italics. χ^2-differences are reported for nested models.

individuals (Lubke et al., 2003a; Hamaker, 2007) as, according to the fit indices used, uniform bias is likely to be present for at least some of the items. Meaningful comparisons between subjects can be considered feasible as long as they refer to differences in the structure of latent intra-individual variation only. The extent and nature of potential uniform bias between individuals could be the subject of subsequent analyses.

ADDRESSING MI OVER TIME

Strict FI over occasions cannot be tested directly, as we confined this illustration to time-invariant models. However, we can investigate whether strict FI over time is violated in a specific sense. We do this by testing for uniform bias of the indicators with respect to a selected time-varying variable X. This can be cast in terms of a main-effect of X on the indicators additionally to the latent variables (Lubke et al., 2003b).

We extend the time-invariant model for a given individual $i = i^*$ to

$$\mathbf{y}_{i^*,\,t} = \boldsymbol{\tau}_{i^*} + \boldsymbol{\Lambda}_{i^*}\boldsymbol{\eta}_{i^*,\,t} + \boldsymbol{\Gamma}_{i^*}\mathbf{x}_{i^*,\,t} + \boldsymbol{\varepsilon}_{i^*,\,t} \quad (8)$$

and

$$\boldsymbol{\eta}_{i^*,\,t} = \boldsymbol{\alpha}_{i^*} + \mathbf{B}_{i^*}\boldsymbol{\eta}_{i^*,\,t-1} + \boldsymbol{\Phi}_{i^*}\mathbf{x}_{i^*,\,t} + \boldsymbol{\zeta}_{i^*,\,t} \quad (9)$$

where $\boldsymbol{x}_{i*,\,t}$ is a $r \times 1$ vector of (fixed) covariates and $\boldsymbol{\Gamma}_{i*}$ and $\boldsymbol{\Phi}_{i*}$ are $p \times r$ and $q \times r$ matrices of regression coefficients. If there is a significant effect of at least one variable in $\boldsymbol{x}_{i*,\,t}$ on at least one of the indicators, measurement invariance over time would be violated, as—returning to Mellenbergh's definition—the distribution of the indicators is dependent on $\boldsymbol{x}_{i*,\,t}$ conditional on the latent variables (Lubke et al., 2003b). However, the absence of uniform bias with respect to $\boldsymbol{x}_{i*,\,t}$ implies neither MI with respect to these variables (which may still introduce non-uniform bias or be associated with varying measurement residual variances), nor MI with respect to other time-varying variables, let alone MI with respect to time.

We focused on the neuroticism marker item "bad tempered" as a mood indicator and potentially biasing variable in subject 7. The results are shown in **Table 3** and the path diagrammatic representation of the corresponding model is displayed in **Figure 3**.

The BIC which is more responsive to parsimony than the AIC (Hamaker et al., 2005) favors the model without direct effect of the mood indicator on all indicators and the agreeableness indicators respectively. Both AIC and χ^2-difference test suggest that uniform bias is present for at least one of the indicators. In a given modeling application one could investigate whether uniform bias can be accounted or controlled for also with respect to other potentially biasing covariates. Ultimately however, one needs to decide whether one is willing to discard other forms of bias over time as unlikely or whether actually a modeling approach that incorporates time-varying parameters is the more valid and more interesting alternative. Fitting the "wrong" model to intra-individual data which could be a measurement-invariant or more generally a time-invariant model, will also affect the quality of between-person comparisons. We briefly outline modeling approaches to time-varying dynamics in the discussion.

DISCUSSION

In this paper, we showed how MI (e.g., Mellenbergh, 1989), if present, may facilitate or, if absent, may complicate the generalizability of modeling results within and between subjects. Tying into the ergodicity debate (e.g., Molenaar, 2004), we clarified the relationship between the concepts of MI and ergodicity in the context of general latent variable modeling as well as in

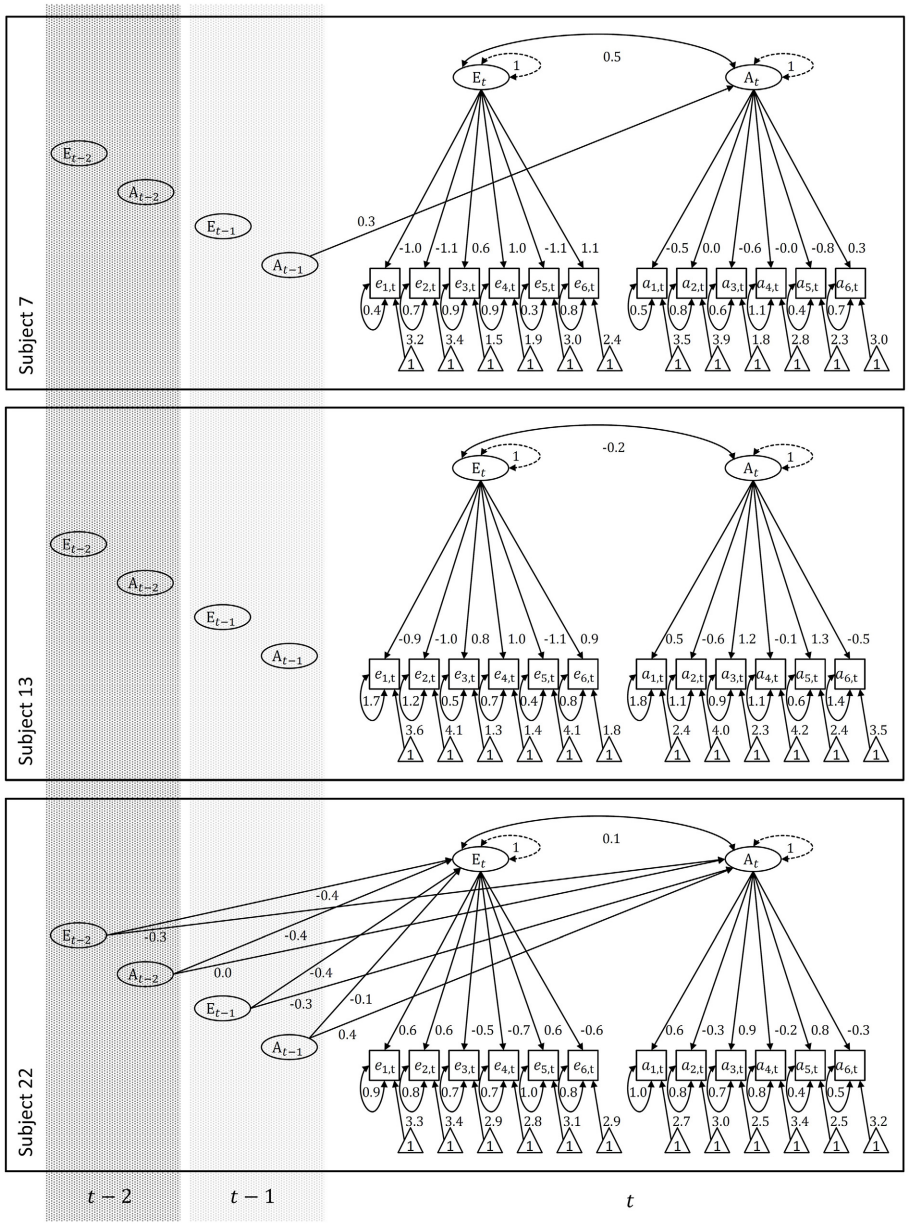


FIGURE 2 | Relatively best fitting models for subjects 7, 13, and 22. Paths fixed to zero are not drawn. Note that these include the regression parameters of the vector eta on the constant, i.e., vector alpha, which are fixed to zero for scaling purposes. Paths fixed to one are dashed. These include the latent residual variances in order to provide a latent metric. Freely estimated paths are drawn in black and parameter point estimates are provided. Items denoted with e are extraversion marker items, whereas items denoted with a are agreeableness marker items. The numerical ordering of the items employed here corresponds to the ordering of the items as given in the data description section. Index *i* is dropped as the models describe single individuals.

a linear multi-subject state-space time series model. We concluded that MI holding simultaneously over time and subjects implies a mode of structural equivalence between the intra- and the inter-individual level of analysis that is distinct from full structural equivalence, i.e., ergodicity. That is, measurement equivalence is a mode of structural equivalence conditional on the latent process. Following common interpretations of measurement invariance, the mode of measurement equivalence could be considered an important condition for integrative latent variable modeling of intra- and inter-individual differences (cf. Ellis and van den Wollenberg, 1993, who stress the importance of local homogeneity in IRT-modeling which is tantamount to measurement equivalence; cf. Millsap, 2011). Using intra-individual time series data from three individuals on daily personality states, we investigated the tenability of MI constraints over subjects and over time. Although strict FI over subjects was absent, the presence of weak FI suggested that between-subject comparisons were feasible with respect to the structure of latent intra-individual variation.

Table 2 | Multi-group models with measurement parameters constrained over groups.

Measurement models	npars	−2LogL	AIC	BIC	χ^2-increase (relative to)	*df*	*p*
Comparison between subjects 7 and 13							
Configural invariance	75	2604	2754	2942			
Weak FI ($\boldsymbol{\Lambda}$ *invariant*)	65	2621	2751	2913			
Strong FI ($\boldsymbol{\Lambda}$, $\boldsymbol{\tau}$ invariant)	55	2797	2907	3044			
Strict FI ($\boldsymbol{\Lambda}$, $\boldsymbol{\tau}$, $\boldsymbol{\Theta}$ invariant)	43	2863	2949	3056	66.087(Strong FI)	12	0.000
Comparison between subjects 7 and 22							
Configural invariance	83	2242	2408	2616			
Weak FI ($\boldsymbol{\Lambda}$ *invariant*)	73	2255	2401	2583			
Strong FI ($\boldsymbol{\Lambda}$, $\boldsymbol{\tau}$ invariant)	63	2474	2600	2757			
Strict FI ($\boldsymbol{\Lambda}$, $\boldsymbol{\tau}$, $\boldsymbol{\Theta}$ invariant)	51	2516	2618	2745	42.156(Strong FI)	12	0.000
Comparison between subjects 13 and 22							
Configural invariance	82	2684	2848	3053			
Weak FI ($\boldsymbol{\Lambda}$ *invariant)*	72	2701	2845	3025			
Strong FI ($\boldsymbol{\Lambda}$, $\boldsymbol{\tau}$ invariant)	62	2787	2911	3066			
Strict FI ($\boldsymbol{\Lambda}$, $\boldsymbol{\tau}$, $\boldsymbol{\Theta}$ invariant)	50	6162	6262	6387	3374.630(Strong FI)	12	0.000

The relatively best fitting model according to AIC and BIC is set in italics. χ^2*-differences are reported for nested models.*

Table 3 | Comparison of models incorporating a potentially biasing variable x for subject 7.

Model	npars	−2LogL	AIC	BIC	χ^2-increase (relative to)	*df*	*p*
$\boldsymbol{y}$, $\boldsymbol{\eta}$ on $\boldsymbol{x}$	52	1010	1114	1244			
$\boldsymbol{\eta}$ on $\boldsymbol{x}$	40	1044	1124	1224	34.250 ($\boldsymbol{y}$, $\boldsymbol{\eta}$ on $\boldsymbol{x}$)	12	0.001
$\boldsymbol{y}$(*a*), $\boldsymbol{\eta}$(*a*) on $\boldsymbol{x}$	45	1034	1124	1237			
$\boldsymbol{\eta}$(*a*) on $\boldsymbol{x}$	39	1049	1127	1225	15.061 ($\boldsymbol{y}$(*a*), $\boldsymbol{\eta}$(*a*) on $\boldsymbol{x}$)	6	0.020

$\boldsymbol{y}$(a) *denotes the agreeableness marker items, and* $\boldsymbol{\eta}$(*a*) *denotes the agreeableness factor. We allowed for direct effects of* $\boldsymbol{x}$ *on the latent variables but did not establish whether these were significant.*

χ^2*-differences are reported for nested models.*

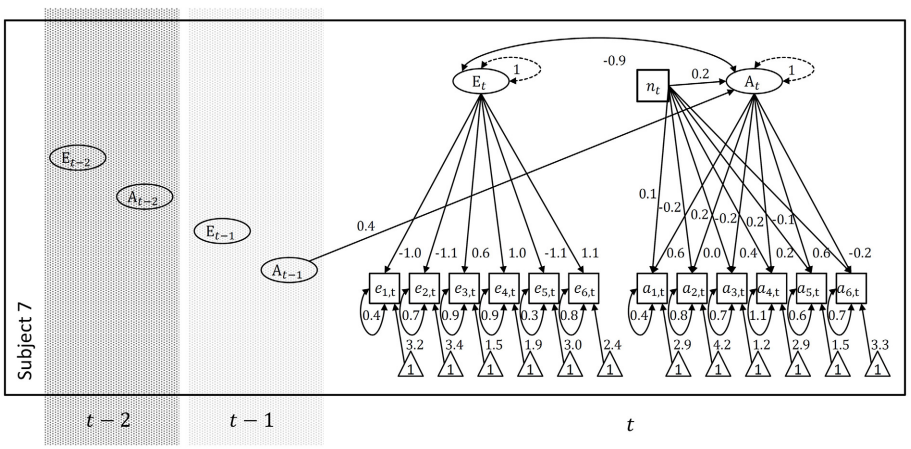


FIGURE 3 | Individual model for subject 7 including the neuroticism marker item "bad tempered" as a potentially biasing (fixed) variable. According to this representation, the neuroticism item possibly affects the agreeableness marker items above the potential effect it has through the agreeableness factor.

We were limited in investigating MI over time due to the time-invariant models we employed. Consequently, we could test for specific MI violations but we did not address unbiasedness with respect to time.

The results of our illustration are in line with a growing body of empirical work investigating potential relationships between the structures of intra- and inter-individual variation and means. So, although we presented measurement equivalence as a less

restrictive mode of equivalence between levels of analysis than full structural equivalence, we acknowledge that even this weaker form of structural equivalence may be overly restrictive. We can therefore only stress that the problem of non-ergodicity must in part be viewed as a measurement problem since the violation of measurement invariance with respect to time and subject is a source of heterogeneity within and between individuals (cf. Nesselroade et al., 2007, 2009; Borsboom et al., 2009). It was the aim of this paper to show that the investigation of measurement related heterogeneity within and between individuals in latent variable modeling qualifies as a problem which is related to but also distinct from the problem of ergodicity.

Regarding a closer examination of measurement related heterogeneity, the presented taxonomy is clearly an abstraction. In practice, the finding of untenable MI constraints is not necessarily the end of an investigation. Modeling application situations falling in the baseline model category and associated problems of measurement variance can be of very different nature. For instance, it may be possible to interpret measurement variance substantively against a given theoretical background (Millsap and Hartog, 1988; Kelderman and Molenaar, 2007). As an example, consider developmental or interventional effects over time, which may manifest as quantitative changes in given parameters, and, more importantly, in changes in the nature or meaning of the psychological entities of interest (Millsap and Hartog, 1988; Molenaar, 2004; Kelderman and Molenaar, 2007; Schmiedek et al., 2009). Also, even if measurement variance is considered a nuisance factor, only a few indicators may display measurement variance. Subsequent analyses may then locate the MI violation in the model and establish whether the number of unbiased indicators is sufficient to proceed with meaningful latent variable modeling, as we have indicated in the illustration (Byrne et al., 1989; Wicherts and Dolan, 2010). Likewise, not all subjects within a sample and not all occasions within a period of time may be affected by measurement variance. It may then be possible to identify intra- or inter-individual variables that explain measurement variance (Mellenbergh, 1989). In the present context, this relates to the concept of conditional equivalence introduced by Voelkle et al. (2014). In a simulation study these authors show that full equivalence between inter- and intra-individual model structures can easily be obscured by incorporating single factors that introduce subject- and time-related heterogeneity, e.g., linear mean trends over time, differences between groups of individuals. Conversely, it might be possible to identify such factors for certain constructs and control for them in order to establish conditional equivalence, that is, equivalence for subgroups of individuals and occasions. In case equivalence is well hidden or absent, one can still explore the various types of less restrictive (unconditional) relationships that may arise between intra-individual and inter-individual model characteristics (cf. Kuppens et al., 2010; Montpetit et al., 2010; Brose et al., 2014).

These approaches to the links between levels of analysis have yet to be utilized to specifically address measurement variance within and between individuals. To further emphasize why these approaches could be both interesting and necessary given measurement related heterogeneity within and between individuals, let us return to the assumptions, upon which MI is predicated. These concern the existence of the latent variables of interest and the appropriateness of the observed variables as indicators. The first premise holds, that the indicators are—although possibly imperfect, i.e., biased—valid in principle (cf. Meredith, 1964, 1993). That is, the indicators are to some extent measuring the variable they were designed to measure (Millsap, 2011) and these psychometric qualities should hold absolutely true or at least hold true for the units of analysis we wish to compare, say, a sample of individuals (Nesselroade et al., 2009). This in turn requires the assumption that the targeted latent variable is indeed given (Mellenbergh, 1989) or a theoretically sensible construct across the selected individuals. As noted by Byrne and Campbell (1999) these premises may be questionable, for instance in applying a measurement instrument in a setting, other than the setting in which it was developed. The setting may be determined by the cultural background of the examinees or the dimension of analysis, e.g., the intra-individual dimension. Hence, a violation of MI with respect to differing setting conditions can be indicative in the following regard. First, it may be that the given test is not valid under some conditions although the latent variable is—on an abstract level—existent or theoretically sensible. The latent variable simply manifests differently under different conditions (e.g., Byrne and Campbell, 1999). Nesselroade et al. (2007, 2009) pointed out that a targeted construct (e.g., athletic performance) may be a sensible choice for comparing different individuals—but may require the use of individual-specific indicators ("How well do you play tennis vs. golf?"). Second, a given test may be invalid under certain conditions because the construct is not conceptually sensible across conditions. To label these two scenarios, Byrne and Campbell (1999) refer to the term *construct bias* as opposed to item bias which indicates that the problem has shifted from an "operational" to a "theoretical" problem (Kelderman and Molenaar, 2007, p. 451). The concept of construct bias seems to be highly interesting when contrasting intra- and inter-individual variation. In the light of increasing empirical evidence in favor of substantive individual specifics (e.g., Hamaker et al., 2005; Brose et al., 2010) it raises the following question: To what extent are traditional psychological constructs (and according measurement instruments) that were derived in a between-subject context applicable to intra-individual differences? This is arguably a philosophical question, which has been addressed intensively by Borsboom et al. (2003, 2009) and by Cervone (2004, 2005). These authors argue that between-subject constructs like extraversion and agreeableness do well in describing inter-individual differences, but are problematic at the level of the individual, where they lack "causal force" (e.g., Cervone, 2004; p. 184). That is, *per se*, they do not map onto specific psychological mechanisms or processes within the individual, and are thus not suitable to feature as explaining factors in a within-subject model of psychological functioning (van der Maas et al., 2006; Borsboom et al., 2009). Borsboom et al. (2009) conjecture that there are "infinitely many ways" (p. 88) to achieve a certain outcome on a standard between-subject dimension. The associated constructs thus may lack coherence from an individual-driven perspective, in that they emerge as abstract aggregates only at the level of the population. However, this pessimistic prospect regarding the meaningful application of inter-individual level constructs to the individual

can be probed empirically. Millsap employs the term *differential item functioning* rather than the term bias to indicate that "the researcher is unable or unwilling to clearly define the targeted attribute" (Millsap, 2011; p. 9). This can be turned into a positive message, namely to explore measurement variance—be it within or between individuals—as a potentially meaningful phenomenon.

An explorative empirical approach to person- and time-related heterogeneity at the level of measurement using the above described strategies and principles can enlighten how measurement instruments that were constructed in the between-subject context function at the within-subject level. This in turn can inform (and be informed by) the elaboration of individual-level concepts and theories (e.g., Cervone, 2005) as well as their implementation in empirical research in terms of operationalizations, measurement devices, and modeling techniques (e.g., Schmiedek et al., 2009). In this sense, it could contribute to building up the theoretical and conceptual foundation that is needed for a true reorientation toward the individual in differential psychology (Molenaar, 2004).

The presented modeling approach has the following limitations, however, that would restrict such an explorative endeavor. First, we based our modeling on the linear, time-invariant Kalman filter and ML estimation which led to time-invariant time series models. Time-varying model parameters can—to some extent—be accommodated using the extended Kalman filter (e.g., Chow et al., 2011; Chow and Zhang, 2013) or a Bayesian approach (e.g., Del Negro and Otrok, 2008). Second, we employed a multi-group approach, i.e., a two-step procedure to address inter-individual differences in intra-individual dynamics. Inter-individual differences in intra-individual model parameters can be quantified and modeled directly using a Bayesian multi-level approach (e.g., Lodewyckx et al., 2011). Note, however, that multi-group modeling is in principle less restrictive than hierarchical modeling. In the present context, it did not impose any restrictions across individuals apart from applying the same modeling framework to each individual's data. That is, within individuals, we assumed continuous, normal variables, at the manifest and latent level, which were linearly related to each other. Our reliance on the linear factor model here is expedient, although we are satisfied linear modeling of 7 point scales is adequate. Generalized linear modeling of intra-individual time series to accommodate discrete indicators is possible (cf. van Rijn et al., 2010), but at present depends on software development. Non-normally distributed continuous indicators (due to nonlinear effects) can be approximated by mixtures of (un-)conditional normal distributions (e.g., Klein and Moosbrugger, 2000). Note that in our case of single-subject models, mixture models return us to time-varying models (Hunter, 2014), which are increasingly discussed in the psychometric literature.

ACKNOWLEDGMENTS

Janne Adolf thanks her colleagues from the "intra-person behavioral dynamics" project at the Max Planck Institute for Human Development for their helpful comments and discussion input. Janne Adolf is a pre-doctoral fellow of the International Max Planck Research School on the Life Course (LIFE, www.imprs-life.mpg.de; participating institutions: MPI for Human Development, Freie Universität Berlin, Humboldt-Universität zu Berlin, University of Michigan, University of Virginia, University of Zurich). Conor V. Dolan is supported by the European Research Council (Genetics of Mental Illness: ERC-230374 awarded to Dorret I. Boomsma).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00883/abstract

REFERENCES

Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Automat. Contr.* 19, 716–723. doi: 10.1109/TAC.1974.1100705

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annu. Rev. Psychol.* 53, 605–634. doi: 10.1146/annurev.psych.53.100901.135239

Borkenau, P., and Ostendorf, F. (1990). Comparing exploratory and confirmatory factor analysis: a study on the 5-factor model of personality. *Pers. Individ. Dif.* 11, 515–524. doi: 10.1016/0191-8869(90)90065-Y

Borkenau, P., and Ostendorf, F. (1998). The big five as states: How useful is the five factor model to describe inraindividual variations over time? *J. Res. Pers.* 32, 202–221. doi: 10.1006/jrpe.1997.2206

Borsboom, D. (2008). Latent variable theory. *Measurement* 6, 25–53. doi: 10.1080/15366360802035497

Borsboom, D., and Dolan, C. V. (2007). Theoretical equivalence, measurement, invariance, and the idiographic filter. *Measurement* 5, 236–263. doi: 10.1080/15366360701765020

Borsboom, D., Kievit,. R. A., Cervone, D., and Hood, B. S. (2009). "The two disciplines of scientific psychology, or: the disunity of psychology as a working hypothesis," in *Developmental Process Methodology in the Social and Developmental Sciences*, eds J. Valsiner, P. C. M. Molenaar, M. C. D. P. Lyra, and N. Chaudary (New York, NY: Springer), 67–97.

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2003). The theoretical status of latent variables. *Psychol. Rev.* 110, 203–219. doi: 10.1037/0033-295x.110.2.203

Brose, A., Schmiedek, S., Lövdén, M., Molenaar, P. C. M., and Lindenberger, U. (2010). Adult age differences in covariation of motivation and working memory performance: contrasting between-person and within-person findings. *Res. Hum. Dev.* 7, 61–78. doi: 10.1080/15427600903578177

Brose, A., Voelkle, M. C., Lövdén, M., Lindenberger, U., and Schmiedek, F. (2014). Differences in the between-person and within-person structures of affect are a matter of degree. *Eur. J. Psychol*. doi: 10.1002/per.1961

Byrne, B. M., and Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: a look beneath the surface. *J. Cross Cult. Psychol.* 30, 555–574. doi: 10.1177/0022022199030005001

Byrne, B. M., Shavelson, R. J., and Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Cattell, R. B. (1952). The three basic factor-analytic designs: their interrelations and derivatives. *Psychol. Bull.* 49, 499–520. doi: 10.1037/h0054245

Cervone, D. (2004). The architecture of personality. *Psychol. Rev.* 111, 183–204. doi: 10.1037/0033-295X.111.1.183

Cervone, D. (2005). Personality architecture: within-person structures and processes. *Annu. Rev. Psychol.* 56, 423–452. doi: 10.1146/annurev.psych.56.091103.070133

Chow, S.-M., Ho, R. M., Hamaker, E. L., and Dolan, C. V. (2010). Equivalences and differences between structural equation modeling and state-space modeling techniques. *Struct. Equ. Model.* 17, 303–332. doi: 10.1080/10705511003661553

Chow, S.-M., and Zhang, G. (2013). Non-linear regime-switching state-space (rsss) models. *Psychometrika* 78, 740–768. doi: 10.1007/S11336-013-9330-8

Chow, S.-M., Zu, J., Shifren, K., and Zhang, G. (2011). Dynamic factor analysis models with time-varying parameters. *Multivariate Behav. Res.* 46, 303–339. doi: 10.1080/00273171.2011.563697

Del Negro, M., and Otrok, C. (2008). "Dynamic factor models with time-varying parameters: measuring changes in international business cycles," in *Federal Reserve Bank of New York Staff Reports, No. 326*. Available online at: http://hdl.handle.net/10419/60779

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7 response categories: A comparison of categorical variable estimators using simulated data. *Br. J. Math. Stat. Psychol.* 47, 309–326. doi: 10.1111/j.2044-8317.1994.tb01039.x

Dolan, C. V. (2009). "Structural equation mixture modeling," in *The Sage Handbook of Quantitative Methods in Psychology*, eds R. E. Millsap and A. Maydeu-Olivares (Thousand Oaks, CA: Sage Publications Ltd.), 568–591. doi: 10.4135/9780857020994.n23

Dolan, C. V. (2010). *MKFM6: Multi-Group, Multi-Subject Stationary Time Series Modeling Based on the Kalman Filter.* Amsterdam: University of Amsterdam.

Dolan, C. V., Roorda, W., and Wicherts, J. M. (2004). Two failures of spearman's hypothesis: the GATB in holland and the JAT in south africa. *Intelligence* 32, 155–173. doi: 10.1016/j.intell.2003.09.001

Durbin, J., and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods*. New York, NY: Oxford University Press.

Ellis, J. L., and van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models: a characterization of the homogeneous monotone IRT model. *Psychometrika* 58, 417–429. doi: 10.1007/bf02294649

Fleeson, W. (2001). Toward a structure- and process-integrated view of personality: traits as density distributions of states. *J. Pers. Soc. Psychol.* 89, 1011–1027. doi: 10.1037/0022-3514.80.6.1011

Graziano, W. G., and Tobin, R. M. (2009). "Agreeableness," in *Handbook of Individual Differences in Social Behavior*, eds M. R. Leary and R. H. Hoyle (New York, NY: Guilford Press), 46–61.

Gregson, R. A. M. (1983). *Time Series in Psychology.* Hillsdale, NJ: Lawrence Erlbaum Associates. Inc.

Hamaker, E. J., Dolan, C. V., and Molenaar, P. C. M. (2005). Statistical modeling of the individual: Rationale and application of multivariate stationary time series analysis. *Multivariate Behav. Res.* 40, 207–233. doi: 10.1207/s15327906mbr4002_3

Hamaker, E. L. (2007). How to inspect fruit. *Measurement* 5, 250–253. doi: 10.1080/15366360701775961

Hamaker, E. L., and Dolan, C. V. (2009). "Idiographic data analysis: quantitative methods - from simple to advanced," in *Dynamic Process Methodology in the Social and Developmental Sciences*, eds J. Valsiner, P. C. M. Molenaar, M. Lyra, and N. Chaudhary (New York, NY: Springer-Verlag), 191–216.

Hamaker, E. L., Nesselroade, J. R., and Molenaar, P. C. M. (2007). The integrated trait-state model. *J. Res. Pers.* 41, 295–315. doi: 10.1016/j.jrp.2006.04.003

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Pers. Soc. Psychol. Bull.* 30, 161–172. doi: 10.1177/0146167203259930

Horn, J. L., and McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Exp. Aging Res.* 18, 117–144. doi: 10.1080/03610739208253916

Hox, J. J. (2002). *Multilevel Analysis: Techniques and Applications*. Mahwah, NJ: Erlbaum.

Hunter, M. D. (2014). Abstract: dynamic mixture modeling of a single simulated case. *Multivariate Behav. Res.* 49, 286–287. doi: 10.1080/00273171.2014.912890

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika* 36, 409–426. doi: 10.1007/BF02291366

Kelderman, H., and Molenaar, P. C. M. (2007). The effect of individual differences in factor loadings on the standard factor model. *Multivariate Behav. Res.* 42, 435–456. doi: 10.1080/00273170701382997

Klein, A., and Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika* 65, 457–474. doi: 10.1007/BF02296338

Kuppens, P., Allen, N. B., and Sheeber, L. B. (2010). Emotional inertia and psychological maladjustment. *Psychol. Sci.* 21, 984–991. doi: 10.1177/0956797610372634

Lodewyckx, T., Tuerlinckx, F., Kuppens, P., Allen, N. B., and Sheeber, L. (2011). A hierarchical state space approach to affective dynamics. *J. Math. Psychol.* 55, 68–83. doi: 10.1016/j.jmp.2010.08.004

Lubke, G. H., Dolan, C. V., Kelderman, H., and Mellenbergh, G. J. (2003a). On the relationship between- and within group differences and measurement invariance in the common factor model. *Intelligence* 31, 543–566. doi: 10.1016/s0160-2896(03)00051-5

Lubke, G. H., Dolan, C. V., Kelderman, H., and Mellenbergh, G. J. (2003b). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *Br. J. Math. Stat. Psychol.* 56, 231–248. doi: 10.1348/000711003770480020

Lubke, G. H., and Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Struct. Equ. Model.* 11, 514–534. doi: 10.1207/s15328007sem1104_2

McCrae, R. R., and John, O. P. (1992). An introduction to the five-factor model and its applications. *J. Pers.* 60, 175–215. doi: 10.1111/j.1467-6494.1992.tb00970.x

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behav. Res.* 29, 223–236. doi: 10.1207/s15327906mbr2903_2

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika* 29, 177–185. doi: 10.1007/BF02289699

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Millsap, R. E., and Hartog, S. B. (1988). Alpha, beta, and gamma change in evaluation research: a structural equation approach. *J. Appl. Psychol.* 73, 574–584. doi: 10.1037/0021-9010.73.3.574

Molenaar, P. C. M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika* 50, 181–202. doi: 10.1007/bf02294246

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement* 2, 201–218. doi: 10.1207/s15366359mea0204_1

Molenaar, P. C. M., and Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Curr. Dir. Psychol.* 18, 112–117. doi: 10.1111/j.1467-8721.2009.01619.x

Molenaar, P. C. M., Huizenga, H. M., and Nesselroade, J. R. (2003). "The relationship between the structure of interindividual and intraindividual variability: a theoretical and empirical vindication of developmental systems theory," in *Understanding Human Development: Dialogues with Lifespan Psychology,* eds U. M. Staudinger and U. Lindenberger (Dordrecht: Kluwer Academic Publishers), 339–360.

Montpetit, M. A., Bergeman, C. S., Deboeck, P. R., Tiberio, S. S., and Boker, S. M. (2010). Resilience-as-process: negative affect, stress, and coupled dynamical systems. *Psychol. Aging* 25, 631–640. doi: 10.1037/a0019268

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika* 54, 557–585. doi: 10.1007/BF02296397

Nesselroade, J. R. (2010). "On an emerging third discipline of scientific psychology," in *Individual Pathways of Change: Statistical Models for Analyzing Learning and Development*, eds P. C. M. Molenaar and K. M. Newell (Washington, DC: American Psychological Association), 209–218.

Nesselroade, J. R., Gerstorf, D., Hardy, S. A., and Ram, N. (2007). Idiographic filters for psychological constructs. *Measurement* 5, 217–235. doi: 10.1080/15366360701741807

Nesselroade, J. R., Ram, N., Gerstorf, D., and Hardy, S. A. (2009). Rejoinder to commentaries on Nesselroade, Gerstorf, Hardy, and Ram. *Measurement* 7, 17–26. doi: 10.1080/15366360802715361

Oud, J. H. L., van den Bercken, J. H., and Essers, R. J. (1990). Longitudinal factor score estimation using the Kalmanfilter. *Appl. Psychol. Meas.* 14, 395–418. doi: 10.1177/014662169001400406

Paulhus, D. L., and Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *J. Pers. Soc. Psychol.* 60, 307–317. doi: 10.1037/0022-3514.60.2.307

Rammstedt, B., and John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K): Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit. *Diagnostica* 51, 195–206. doi: 10.1026/0012-1924.51.4.195

Raykov, T., Marcoulides, G. A., and Li, C. H. (2012). Measurement invariance for latent constructs in multiple populations: a critical view and refocus. *Educ. Psychol. Meas.* 72, 954–974. doi: 10.1177/0013164412441607

Schmiedek, F., Lövdén, M., and Lindenberger, U. (2009). On the relation of mean reaction time and intraindividual reaction time variability. *Psychol. Aging* 24, 841–857. doi: 10.1037/a0017799

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Shumway, R. S., and Stoffer, D. S. (2011). *Time Series Analysis and its Applications: With R Examples*. New York, NY: Springer. doi: 10.1007/978-1-4419-7865-3

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., and Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychol. Rev.* 113, 842–861. doi: 10.1037/0033-295X.113.4.842

van der Sluis, S., Posthuma, D., Dolan, C. V., de Geus, E. J. C., Colom, R., and Boomsma, D. I. (2006). Sex differences on the dutch WAIS-III. *Intelligence* 34, 273–289. doi: 10.1016/j.intell.2005.08.002

van Rijn, P., Dolan, C. V., and Molenaar, P. C. M. (2010). "State space methods for item response modeling of multisubject time series," in *Individual Pathways of Change: Statistical Models for Analyzing Learning and Development*, eds P. C. M. Molenaar and K. M. Newell (Washington, DC: American Psychological Association), 125–135.

Voelkle, M. C., Brose, A., Schmiedek, F., and Lindenberger, U. (2014). Toward a unified framework for the study of between-person and within-person structures: building a bridge between two research paradigms. *Multivariate Behav. Res.* 49, 193–213. doi: 10.1080/00273171.2014.889593

Wicherts, J. M., and Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: an illustration using IQ test performance of minorities. *Educ. Meas.* 29, 39–47. doi: 10.1111/j.1745-3992.2010.00182.x

Conflict of Interest Statement: The Review Editor Ellen Hamaker declares that, despite having been supervisor of author Noémi K. Schuurman, who is also affiliated with the same institution and whom they collaborated with, the review process was handled objectively and no conflict of interest exists. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2014; accepted: 24 July 2014; published online: 19 September 2014.

Citation: Adolf J, Schuurman NK, Borkenau P, Borsboom D and 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. doi: 10.3389/fpsyg.2014.00883*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 29 September 2014
doi: 10.3389/fpsyg.2014.01087

Measurement bias detection through Bayesian factor analysis

M. T. Barendse[1]*, C. J. Albers[1], F. J. Oort[2] and M. E. Timmerman[1]

[1] Psychometrics and Statistics, Heymans Institute for Psychological Research, University of Groningen, Groningen, Netherlands
[2] Department of Education, Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, Netherlands

Edited by:
Rens Van De Schoot, Utrecht University, Netherlands

Reviewed by:
Megan Welsh, University of Connecticut, USA
Yanyan Sheng, Southern Illinois University, USA
Sarah Depaoli, University of California, Merced, USA

***Correspondence:**
M. T. Barendse, Psychometrics and Statistics, Heymans Institute for Psychological Research, University of Groningen, Grote Kruisstraat 2/1, 9712TS Groningen, Netherlands
c/o e-mail: c.j.albers@rug.nl

Measurement bias has been defined as a violation of measurement invariance. Potential violators—variables that possibly violate measurement invariance—can be investigated through restricted factor analysis (RFA). The purpose of the present paper is to investigate a Bayesian approach to estimate RFA models with interaction effects, in order to detect uniform and nonuniform measurement bias. Because modeling nonuniform bias requires an interaction term, it is more complicated than modeling uniform bias. The Bayesian approach seems especially suited for such complex models. In a simulation study we vary the type of bias (uniform, nonuniform), the type of violator (observed continuous, observed dichotomous, latent continuous), and the correlation between the trait and the violator (0.0, 0.5). For each condition, 100 sets of data are generated and analyzed. We examine the accuracy of the parameter estimates and the performance of two bias detection procedures, based on the DIC fit statistic, in Bayesian RFA. Results show that the accuracy of the estimated parameters is satisfactory. Bias detection rates are high in all conditions with an observed violator, and still satisfactory in all other conditions.

Keywords: Bayesian structural equation modeling, measurement invariance, uniform bias, nonuniform bias, interaction effects

1. INTRODUCTION

Measurement bias research examines whether different respondents show differences in response behavior to test items. In the presence of measurement bias, systematic differences between observed test scores do not validly represent differences in the trait(s) that the test is supposed to measure. Measurement bias is formally defined as a violation of measurement invariance (Oort, 1992, after Mellenbergh, 1989). A test is measurement invariant with respect to V, if the following conditional independence holds:

$$f_1(X|T = t, V = v) = f_2(X|T = t), \quad (1)$$

where X is a set of observed variables, T the trait(s) of interest measured by X, and V a set of variable(s) other than T, which possibly violates measurement invariance; function f_1 is the conditional distribution of X given values of t and v, and f_2 is the conditional distribution of X given t. If conditional independence does not hold (i.e., $f_1 \neq f_2$), the measurement of T by X is said to be biased with respect to V. This is a general definition of measurement bias in the sense that T and V may be measured on any measurement level (i.e., nominal, ordinal, interval, or ratio), and their mutual relationships may be linear or non-linear. In addition, the violator V may be observed or latent.

Structural equation modeling (SEM) offers a flexible framework to test for measurement bias. If the violator is an observed discrete variable, e.g., indicating group membership, measurement bias is typically investigated through multigroup factor analysis (MGFA; Meredith, 1993). In MGFA, differences across groups in intercepts indicate uniform bias (i.e., the size of bias is constant across the trait levels) and differences across groups in factor loadings indicate nonuniform bias (i.e., the size of bias varies with the trait levels). Because MGFA requires an observed discrete violator, its use is rather restricted. In the case of a continuous violator, MGFA is sometimes applied using the categorized version of the violator. However, this practice is to be discouraged, because of the known negative consequences of categorizing variables (e.g., MacCallum et al., 2002; Barendse et al., 2012), and because an attractive alternative is available. This very generally applicable alternative is restricted factor analysis (RFA; Oort, 1992, 1998). The advantages of RFA over MGFA are that RFA can assess measurement bias with respect to any kind of violator (i.e., continuous or discrete, observed, or latent) and with respect to multiple violators simultaneously.

In the linear model associated with RFA, the model for x_i, a vector with the observed scores for subject i on J variables X, with a single violator and a single latent trait is defined as

$$x_i = u + at_i + bv_i + ct_iv_i + de_i, \quad (2)$$

where u is a vector of intercepts for the J observed variables, t_i and v_i are the scores of subject i on the latent trait T and the potential violator V, respectively, e_i is a vector of subject i's scores on the standard residual factors E, and a, b, c, and d are vectors of regression coefficients; the elements of a and d^2 are typically denoted as the loadings and residual variances, respectively, and b and c express possible bias. In case V is a categorical variable, dummy variables are used for V. If the relationships between the potential violator and the observed variables are

entirely explained by the indirect relationships through the latent trait, then the observed variables are unbiased with respect to the possible violator. A non-zero element in b indicates uniform bias, and a non-zero element in c indicates nonuniform bias. Uniform bias can thus be investigated by testing the direct effects of a violator on the observed variables (Oort, 1992, 1998). Non-uniform bias can be investigated by testing the direct effect of the product of the latent trait and the violator on the observed variables (see Barendse et al., 2010, 2012), either by using latent moderated structures (Klein and Moosbrugger, 2000) or by using a random slope parametrization (Muthén and Asparouhov, 2003).

The RFA method is similar to the multiple indicator multiple cause (MIMIC) method as described by Muthén (1989). They differ in that in the MIMIC model the violator has a causal effect on the latent trait, whereas in RFA the two are correlated.

So far, the vast majority of the literature on testing measurement bias concerns frequentist methods. Alternatively, a Bayesian approach could be used, thereby offering the general advantage that prior knowledge can be incorporated in the analysis. Recently, the first steps were taken toward a Bayesian approach in this context. A Bayesian MGA has been shown to properly detect bias (Lee, 2007). Further, Muthén and Asparouhov (2012) motivate that the Bayesian approach is more suitable to reflect substantive theories, because it allows for an approximate parameter specification, rather than an exact one. As Muthén and Asparouhov (2013) show, a Bayesian MGFA thus allows for approximate measurement invariance testing. Because a Bayesian MGFA is still restricted to cases with a single observed discrete violator, we consider a Bayesian RFA method here. This method is appealing, because it shares the general advantages of a Bayesian approach, while being applicable to assess measurement bias with respect to multiple violators simultaneously, and of any kind (i.e., continuous or discrete, observed or latent).

The purpose of the present paper is to examine the performance of the Bayesian approach to estimate RFA models with interaction effects, in order to detect uniform and nonuniform measurement bias. An additional advantage of Bayesian RFA is that it handles the estimation of the interaction term easier than frequentist (maximum likelihood) RFA. In a simulation study, we will examine the accuracy of the parameter estimates in the Bayesian RFA, and we will compare the performance of two bias detection procedures.

2. METHODS

Measurement bias in simulated data will be investigated with a Bayesian version of the RFA method. In the data generation, we vary the type of bias (none, only uniform, only nonuniform, both uniform and nonuniform), the type of the continuous violator (observed, latent), and the correlation between the trait and the violator ($\rho(T,V) = 0.0, 0.5$). In a fully crossed design with 100 replications for each condition, this yields $4 \times 2 \times 2 \times 100 = 1600$ simulated datasets. We additionally introduce a dichotomized violator by performing a median split on the observed continuous violator, and thus analyze 2400 datasets in total. Each data set is analyzed using two different bias detection procedures (to be explained in Section 2.4). The accuracy and efficiency of the parameter estimates is assessed. The performance of the two bias detection procedures is evaluated by examining the proportions of true and false positives.

2.1. DATA GENERATION

Each data set consists of the observed scores of 500 subjects on 6 items with continuous response scales, and is generated according to the linear model in Equation 2. We draw subject scores t, v, and e from a multivariate standard normal distribution with an identity covariance matrix in the condition with $\rho = 0.0$; in the condition with $\rho = 0.5$, the element in the covariance matrix associated with t and v is set to 0.500. We have chosen the value of $\rho = 0.0$ as we presume that an absence of linear dependency is the easiest condition in this respect; further we have chosen the value of $\rho = 0.5$, as its corresponds to a "large correlation," according to Cohen's rules of thumb (Cohen, 1988). The intercepts u are set at zero, and the regression coefficients a and d are set at 1.000, for all items. Bias is introduced in the first item only, in such a way that the amount of bias is in line with other bias detection studies (e.g., Oort, 1998). That is, we set parameter $b = 0.400$ to obtain uniform bias and parameter $c = 0.400$ to obtain nonuniform bias—the remaining elements of b and c are fixed at zero. **Table 1** gives an overview of the chosen parameter values for the first item. With these values, if T and V are uncorrelated, the expected percentage of total observed item variance due to the bias is approximately 7% in conditions with only uniform or nonuniform bias and approximately 14% in conditions with both uniform and nonuniform bias. If T and V are correlated, these percentages are 6% (in case of uniform bias), 7% (nonuniform bias), and 13% (both uniform and nonuniform bias).

The violator can either be a continuous latent, a continuous observed or a dichotomous observed variable. In conditions with a continuous latent violator, we introduce three observed variables indicative of the latent violator, which follow a linear factor model. We draw the scores on the latent violator and the residuals independently from a standard normal distribution, and use

Table 1 | Parameter values for 4 (type of bias) × 2 (correlation between trait and violator) = 8 data generation conditions.

	Unstandardized values of Item 1 parameters				
	a	***b***	***c***	***d***	$\sigma^2(X)$
$\rho(T,V) = 0.0$					
No bias	1.000	0.000	0.000	1.000	2.000
Uniform	1.000	0.400	0.000	1.000	2.160
Nonuniform	1.000	0.000	0.400	1.000	2.160
Both	1.000	0.400	0.400	1.000	2.320
$\rho(T,V) = 0.5$					
No bias	1.000	0.000	0.000	1.000	2.000
Uniform	1.000	0.400	0.000	1.000	2.560
Nonuniform	1.000	0.000	0.400	1.000	2.200
Both	1.000	0.400	0.400	1.000	2.760

$u = 0$, $\mu(T) = \mu(V) = \mu(E) = 0$, $\sigma^2(T) = \sigma^2(V) = \sigma^2(E) = 1$; All values pertain to the parameters of Item 1, which is biased in all conditions with bias; parameters of all other items have $a = 1$, $b = 0$, $c = 0$, and $d = 1$ in all conditions. See Appendix 1 in Supplementary Material for the computation of $\sigma^2(X)$.

factor loadings equal to one. In conditions with a continuous observed violator, we draw V from a standard normal distribution. In conditions with a dichotomous observed violator, we perform a median split on the continuous observed violator and conveniently choose $V = -1$ for one group and $V = 1$ for the other group to model the interaction effects.

2.2. BAYESIAN STRUCTURAL EQUATION MODELING AND BIAS DETECTION

Bayesian SEM to detect bias is embedded in Bayesian theory and the associated computational procedures. Bayesian theory combines prior information about the distributions of parameters (called the prior distributions) and the distributions of the data under any SEM model (M). Let θ denote a vector of unknown parameters that are considered to be random. As the observed data and the parameters are random, we model the joint probability (called the posterior distribution) as a function of the conditional distribution of the data given the parameters $p(X|\theta, M)$ and the prior distribution of the parameters $p(\theta)$. More formally this is defined in Bayes' rule:

$$p(X, \theta|M) = \frac{p(X|\theta, M)p(\theta)}{p(X)}, \tag{3}$$

where $p(X)$ normalizes the conditional distribution. As normalizing does not involve any model parameters, Equation 3 can be rewritten as

$$p(X, \theta|M) \propto p(X|\theta, M)p(\theta). \tag{4}$$

Equations 3 and 4 show that the posterior density function includes sample information and prior information. If the prior distribution of θ is so-called uninformative, the posterior density function is proportional to the log-likelihood function. Ideally, a closed form solution of the posterior can be obtained via integration. In practice, one simulates a sufficiently large number of observations from the posterior distribution with Markov Chain Monte Carlo sampling to approximate statistics such as the mean or mode of parameters. Tanner and Wong (1987) introduced the idea to analyze latent variables in a Bayesian context, which is particularly useful for SEM. Latent variables are then treated as hypothetical missing data and the posterior distribution is analyzed on the basis of the complete data.

2.3. BAYESIAN MODEL SELECTION

In Bayesian bias detection we aim at identifying the biased item(s) and the nature of the bias. We therefore compare competing models (i.e., models with and without parameters to account for bias) and select the best fitting model using the deviance information criterion (DIC; see Spiegelhalter et al., 2002). The DIC is a measure of model fit that penalizes for complexity. Under a competing model M_k, the DIC is defined as

$$DIC_k = -\frac{2}{L}\sum_{l=1}^{L} \log p(Y|\theta_k^{(l)}, M_k) + 2d_k \tag{5}$$

where θ_k is a vector of unknown parameters of dimension d_k, and $\{\theta^{(l)} : l = 1, \ldots, L\}$ is a sample of observations simulated from the posterior distribution. The model with the smallest DIC has the highest chance to predict a replicate data set.

Lee (2007) already concluded that a very small difference in DIC values of competing models could be misleading. Also, Lunn et al. (2009) outlines a variety of reasons that could distort the DIC values. We therefore compare our reference model—to be defined later—with competing models and apply two different cut-off values, namely a strict cut-off and a liberal cut-off, to be defined later.

2.4. MEASUREMENT BIAS DETECTION

In a model accounting for bias, we include a direct effect of the violator on the item score to account for uniform bias and a direct effect of the product of the trait and the violator on the item score to account for nonuniform bias. We consider three types of violators (latent continuous, observed continuous, and observed dichotomous), and therefore define three related Bayesian RFA models to model both uniform and nonuniform bias. A bias detection model with respect to a continuous latent violator is graphically displayed in **Figure 1**. To evaluate the approach, we will examine the accuracy and efficiency of the parameter estimates, and the performance in detecting bias with two bias detection procedures.

2.4.1. Parameter estimates

To evaluate the accuracy and the efficiency of the parameter estimates expressing uniform and nonuniform bias, we estimate for each simulated data set the model according to Equation 2; herewith, we fix the elements of b and c associated with items 2–6 (which are non-biased) at zero. Of each converged estimated model, we consider for the first item the posterior distribution of b (indicating uniform bias) and the posterior distribution of c (indicating nonuniform bias). For each posterior distribution, we compute its mean (denoted as $\overline{\theta_b}$ for b and $\overline{\theta_c}$ for c) and the Monte Carlo standard error (denoted as $\text{se}(\theta_b)$ and $\text{se}(\theta_c)$, respectively) using the time series standard error as implemented in CODA (Roberts, 1996; Plummer et al., 2006).

As the accuracy measure, we compute for each condition the estimation bias, as the average of the means of the posterior distribution minus the chosen population values [i.e., $(\text{m}(\overline{\theta_b}) - b)$ and $(\text{m}(\overline{\theta_c}) - c)$]. (Throughout this paper, we shall use the notation

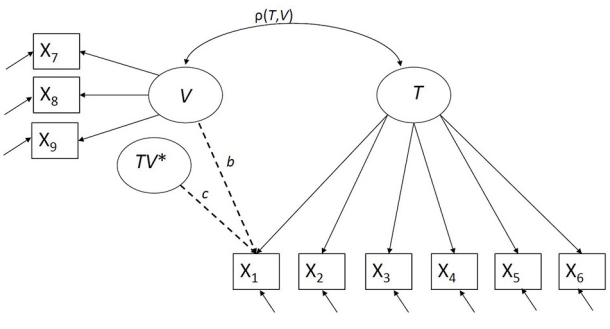


FIGURE 1 | Bias detection with respect to a continuous latent violator *V*, *TV indicates the interaction between *T* and *V*.**

m(·) to denote a mean.) As the efficiency measure, we compute for each condition the standard deviation of the means of the posterior distribution (i.e., sd($\overline{\theta_b}$) and sd($\overline{\theta_c}$)). To evaluate the Monte Carlo sampling accuracy, we compute the mean of the time-series standard errors across the replicates [i.e., m(se(θ_b)) and m(se(θ_c))].

2.4.2. Two procedures to detect bias

In the *single run procedure*, we consider for each of the j ($j = 1, \ldots, 6$) items indicative of the trait, a reference model and compare that model to five (($j' = 1, \ldots, 6$), $j' \neq j$) competing models. Both the reference and competing models have parameters accounting for both uniform and nonuniform bias. For each item j, we consider the model with bias parameters for that item as the reference model. We compare each reference model to each of the five competing models (($j' = 1, \ldots, 6$), $j' \neq j$), by considering each of five DIC differences, as the DIC value of the reference model minus the DIC value of the competing models. With the strict cut-off value, the item with bias parameters in the reference model is indicated as biased if the associated DIC difference is negative. With the liberal cut-off value, the DIC difference should be smaller than -10, to flag item j as biased. The value of 10 was chosen, as it is thought to reflect a substantial difference in model fit (MRC Biostatistics Unit, 2006).

In the *model difference procedure*, we consider for each of the j ($j = 1, \ldots, 6$) items indicative of the trait, a reference model with bias parameters for item j, and compare that to a nested model, namely without any bias parameters. We apply a strict cut-off value, considering item j to be biased whenever the DIC of the reference model is lower than that of the nested model. With the liberal cut-off value, item j is indicated as biased when the value of the DIC of the reference model of item j is at least 10 lower than that of the nested model.

For each of the two procedures, we calculate proportions of true and false positives. A true positive is a biased item that is correctly detected; a false positive is an unbiased item that is incorrectly detected as biased. Proportions of true positives and false positives are interpreted similar to power and Type I error, respectively.

2.5. ANALYSIS

The Bayesian RFA was implemented in R (version 3.02; R Core Team, 2013), using the packages R2OpenBUGS, BRugs, CODA (Plummer et al., 2006) and in BUGS (version 3.2.2; Lunn et al., 2009). All models are fitted to raw data. To estimate non-linear effects, we employ the approach described by Lee (2007), which partitions the latent variables into a linear and a non-linear part with appropriate identification conditions. BUGS uses the Gibbs sampler and the Metropolis-Hastings algorithm for efficient estimation of the Bayesian RFA.

Measurement bias is detected with respect to a continuous latent, a continuous observed and a dichotomous observed violator. Where possible, we use conjugate priors. Conjugate priors are such that the posterior distribution is of the same family as the prior, usually lowering the computational demands of the algorithms drastically. The conjugate priors that we use for each of these models are the normal (for the unknown mean), gamma (for the variance), and the inverse Wishart (for the correlation between latent variables) distributions, as these distributions lead to good results in Bayesian SEM (Lee, 2007; Lee and Song, 2012). We use informative priors that are based on the chosen population values (see **Table 1**). The software cannot work with an observed violator directly. As a workaround, we introduce the violator V as a latent variable with a variance very close to zero, thus making it "practically observed." In conditions with a dichotomous observed violator, we use a prior with a strongly peaked hyperprior for the variance of the violator. Details about the prior elicitation are provided in Appendix 2 in Supplementary Material. All scripts used in this paper are available from http://www.casperalbers.nl.

The number of iterations has been decided upon using the Raftery and Lewis diagnostic (Raftery and Lewis, 1992, 1995). We used 4000 iterations for the burn-in phase and 8000 iterations for the model estimation. For models with a latent violator we used 21000 iterations for the model estimation, as convergence appeared to be slower in these instances in a small pilot study. For each model, we simulated three chains with different initial values for each of the parameters. To decide on the convergence we inspect the Gelman and Rubin's convergence diagnostic (Gelman and Rubin, 1992). This diagnostic compares the within-chain and between-chain variance and a value above 1.1 is an indication of lack of convergence (Gelman et al., 2013). Additionally, we inspected the Geweke convergence diagnostic (Geweke, 1992), which is based on a test for equality of first and the last part of a single Markov chain with the difference between sample means devided by the standard errors expressed in Z-scores.

In addition to the analysis described, we also perform a sensitivity analysis in conditions with a correlated trait and violator (i.e., $\rho(T,V) = 0.5$), and both uniform and nonuniform bias (i.e., Conditions 8, 16, and 24, see **Table 2**), to examine to what extent our choice of the priors influences the parameter estimates. To study the impact of the prior inputs in the Bayesian method, we consider the priors associated with parameters in the absence of bias as inaccurate priors for the bias parameters.

3. RESULTS

After applying the Bayesian RFA to each of the 2400 data sets, we find that the algorithm does not always converge, as indicated by a value exceeding one on Gelman and Rubin's convergence diagnostic (Gelman and Rubin, 1992). Geweke's convergence diagnostic (Geweke, 1992) is much more conservative: it has values larger than the standard threshold value of 2 for at least one chain (out of three) in the vast majority of the simulated data sets, in all conditions. We therefore report convergence according to the Gelman-Rubin diagnostic.

As shown in **Tables 2–4**, we encounter convergence problems especially in conditions that contain uniform bias and a latent violator without a correlation between the trait and the violator and in conditions with an observed violator with a correlation between the trait and the violator.

Non-convergence results are not further analyzed and ignored when assessing the parameter estimates and detecting bias.

Table 2 | Accuracy and efficiency in the Bayesian RFA.

	Bias	Cond.	Conv.	*b*			*c*		
				$m(\overline{\theta_b}) - b$	$sd(\overline{\theta_b})$	$m(se(\theta_b))$	$m(\overline{\theta_c}) - c$	$sd(\overline{\theta_c})$	$m(se(\theta_c))$
CONTINUOUS LATENT VIOLATOR									
$\rho(T,V) = 0.0$	No bias	1	0.95	0.010	0.066	0.001	−0.005	0.073	0.001
	Uniform	2	0.59	0.009	0.062	0.001	0.000	0.078	0.001
	Nonuniform	3	0.93	0.007	0.065	0.001	0.035	0.079	0.001
	Both	4	0.65	0.024	0.065	0.001	0.031	0.087	0.001
$\rho(T,V) = 0.5$	No bias	5	0.97	−0.011	0.081	0.001	−0.004	0.059	0.001
	Uniform	6	0.93	0.005	0.082	0.001	0.000	0.059	0.001
	Nonuniform	7	0.96	−0.014	0.083	0.001	0.021	0.072	0.001
	Both	8	0.92	0.002	0.083	0.001	0.022	0.070	0.001
CONTINUOUS OBSERVED VIOLATOR									
$\rho(T,V) = 0.0$	No bias	9	1.00	0.006	0.054	0.001	−0.002	0.054	0.001
	Uniform	10	0.91	−0.003	0.047	0.001	0.005	0.057	0.001
	Nonuniform	11	0.92	0.000	0.056	0.001	0.008	0.057	0.001
	Both	12	0.92	−0.008	0.059	0.001	0.015	0.051	0.001
$\rho(T,V) = 0.5$	No bias	13	0.80	−0.008	0.054	0.001	−0.006	0.047	0.001
	Uniform	14	0.55	−0.006	0.068	0.001	0.008	0.053	0.001
	Nonuniform	15	0.75	−0.002	0.068	0.001	0.013	0.055	0.001
	Both	16	0.52	−0.004	0.063	0.001	0.006	0.046	0.001
DICHOTOMIZED OBSERVED VIOLATOR (AFTER MEDIAN SPLIT OF THE CONTINUOUS OBSERVED VIOLATOR)									
$\rho(T,V) = 0.0$	No bias	17	1.00	−0.002	0.041	0.001	−0.004	0.054	0.001
	Uniform	18	0.99	−0.089	0.046	0.001	0.003	0.055	0.001
	Nonuniform	19	1.00	0.004	0.051	0.001	−0.071	0.063	0.001
	Both	20	0.98	0.087	0.063	0.001	−0.062	0.058	0.001
$\rho(T,V) = 0.5$	No bias	21	0.94	−0.009	0.051	0.001	0.006	0.056	0.001
	Uniform	22	0.71	−0.112	0.056	0.001	−0.002	0.051	0.001
	Nonuniform	23	0.93	−0.004	0.055	0.001	−0.022	0.071	0.001
	Both	24	0.68	−0.131	0.054	0.001	0.021	0.067	0.001

Cond., Condition; Conv., proportion of converged solutions (of 100 replicates); All summary measures of the parameter estimates are calculated over the converged solutions only.

3.1. PARAMETER ESTIMATES

Table 2 gives the measures of accuracy and efficiency of the estimated parameters that are associated with the parameters that express uniform (i.e., parameter b) and nonuniform (i.e., parameter c) bias in the first item: the estimation bias (i.e., $(m(\overline{\theta_b}) - b)$ and $(m(\overline{\theta_c}) - c)$, the efficiency (i.e., $sd(\overline{\theta_b})$ and $sd(\overline{\theta_c})$), and the Monte Carlo accuracy (i.e., $m(se(\theta_b))$ and $m(se(\theta_c))$).

As can be seen in **Table 2**, the estimation bias appears rather low in the conditions with a continuous latent violator, both for the parameter expressing uniform bias (i.e., b), and nonuniform bias (i.e., c), with a maximum observed estimation bias across all conditions of 0.035. The conditions with a continuous observed violator show a similar pattern, with the largest estimation bias being 0.015.

In the conditions with a dichotomized observed violator, we observe relatively large estimation bias for, firstly, the parameter expressing uniform bias in those conditions that include uniform bias (with a maximum absolute estimation bias of 0.131) and, secondly, but to a lesser extent, the parameter expressing nonuniform bias in those conditions that include nonuniform bias (with a maximum absolute estimation bias of 0.071). With a dichotomized violator, the parameters that represent bias are underestimated.

Across all conditions, the efficiency of the parameters related to uniform and nonuniform bias in the Bayesian RFA is reasonably good, as indicated by the small values of the efficiency parameters ($sd(\overline{\theta_b})$ and $sd(\overline{\theta_c})$) (ranging from from 0.041 to 0.087). We further note that the means of time-series standard errors are small. We therefore conclude that the Monte Carlo accuracy is high.

3.1.1. Single run procedure to detect bias

Table 3 gives the single run procedure results; the convergence rates, the quantile (i.e., 5, 50, and 95) values of the DIC difference between the reference model and the competing model with the most deviating DIC value, and the proportions of true positives and false positives at the strict and the liberal DIC cut-off values. The convergence rates in the single run procedure show considerable variability across conditions (ranging from 0.07 to

Table 3 | Bias detection with the single run procedure.

		Cond.	Conv.	Biased items[a]					Unbiased items[b]				
				Δ DIC			TP		Δ DIC			FP	
				Q_{05}	Q_{50}	Q_{95}	Strict	Liberal	Q_{05}	Q_{50}	Q_{95}	Strict	Liberal
CONTINUOUS LATENT VIOLATOR													
$\rho(T,V) = 0.0$	No bias	1	0.73	–	–	–	–	–	−10	0	0	0.304	0.014
	Uniform	2	0.45	−100	−70	−42	1.000	1.000	−20	0	0	0.387	0.093
	Nonuniform	3	0.67	−110	−70	−40	1.000	1.000	−20	0	0	0.382	0.063
	Both	4	0.52	−185	−125	−86	1.000	1.000	−20	0	0	0.423	0.096
$\rho(T,V) = 0.5$	No bias	5	0.75	–	–	–	–	–	−10	0	0	0.282	0.020
	Uniform	6	0.74	−80	−50	−27	1.000	1.000	−20	0	0	0.432	0.081
	Nonuniform	7	0.78	−140	−85	−40	1.000	1.000	−20	0	0	0.426	0.079
	Both	8	0.75	−193	−130	−70	1.000	1.000	−20	−10	0	0.528	0.157
CONTINUOUS OBSERVED VIOLATOR													
$\rho(T,V) = 0.0$	No bias	9	0.96	–	–	–	–	–	−8	−2	0	0.776	0.016
	Uniform	10	0.88	−103	−71	−50	1.000	1.000	−13	−3	0	0.748	0.093
	Nonuniform	11	0.89	−110	−71	−44	1.000	1.000	−12	−3	0	0.742	0.079
	Both	12	0.88	−175	−138	−93	1.000	1.000	−14	−4	0	0.764	0.143
$\rho(T,V) = 0.5$	No bias	13	0.28	–	–	–	–	–	−8	−2	0	0.756	0.030
	Uniform	14	0.07	−68	−52	−34	1.000	1.000	−14	−2	0	0.657	0.171
	Nonuniform	15	0.18	−106	−77	−46	1.000	1.000	−12	−3	0	0.722	0.100
	Both	16	0.08	−147	−124	−96	1.000	1.000	−18	−7	0	0.750	0.300
DICHOTOMIZED OBSERVED VIOLATOR													
$\rho(T,V) = 0.0$	No bias	17	0.99	–	–	–	–	–	−9	−2	0	0.791	0.022
	Uniform	18	0.99	−70	−45	−28	1.000	1.000	−12	−2	0	0.739	0.089
	Nonuniform	19	0.99	−73	−45	−22	1.000	1.000	−11	−3	0	0.737	0.057
	Both	20	0.96	−119	−88	−55	1.000	1.000	−13	−3	0	0.760	0.104
$\rho(T,V) = 0.5$	No bias	21	0.69	–	–	–	–	–	−8	−2	0	0.775	0.017
	Uniform	22	0.38	−56	−28	−14	1.000	0.974	−11	−3	0	0.758	0.058
	Nonuniform	23	0.62	−78	−50	26	1.000	1.000	−11	−3	0	0.745	0.055
	Both	24	0.32	−118	−84	−54	1.000	1.000	−16	−4	0	0.756	0.131

Cond., Condition; Conv., proportion of converged solutions; Δ DIC denotes the difference in DIC between the reference model and the competing model; [a] Quantile DIC difference values (i.e., 05, 50, 95), and proportions of true positives (TP) are calculated over the converged solutions [of 1 (biased item) × 100 (replicates) = 100 solutions]; [b] Quantile DIC values (i.e., 05, 50, 95), and proportions of false positives (FP) are calculated over the converged solutions, which are 6 (non-biased items) × 100 (replicates) = 600 solutions in Conditions 1, 5, 9, 13, 17 and 21, and 5 (non-biased items) × 100 (replicates) = 500 solutions in all other conditions.

0.99), with particular low values for the models with a continuous latent violator with a substantially correlated latent trait and violator.

As can be seen in **Table 3**, the proportions of true positives (i.e., indicating the bias whenever it is present) are 1.000 in all conditions with the strict criterion, and ranges from 0.974 to 1.000 with the liberal criterion. Thus, both criteria are very successful in detecting the bias. The quantiles of the DIC difference values give an indication of the power to identify items with bias. These DIC difference values are highly negative in conditions with both uniform and nonuniform bias, but also substantial in conditions with only uniform or nonuniform bias. In conditions with a dichotomous observed violator, we observe smaller negative DIC difference values, suggesting a lower power.

The proportions of false positives with the strict cut-off value are very high (ranging from 0.304 to 0.791). With a liberal cut-off value, the proportions of false positives were reasonably low (from 0.014, with a maximum of 0.300); they appear somewhat higher in conditions with both uniform and nonuniform bias and a correlated trait and violator. Considering the performance in terms of both true positives and false positives, the liberal cut-off value seems best suited for bias detection with the single run procedure.

3.1.2. Model difference procedure to detect bias

Table 4 shows the results of the model difference procedure: the convergence proportions, the quantile (i.e., 5, 50, and 95) values of the DIC difference between the reference model (i.e., with parameters to represent bias) and the nested model (i.e., without parameters to represent bias), and the proportions of true positives and false positives at the strict and the liberal cut-off values. The convergence rates show considerable variability across conditions (ranging from 0.49 to 1.00). Overall, the convergence rate is higher in the model difference procedure than in the single run procedure, because the former requires only two, and the latter $J = 6$ models to be estimated.

Table 4 | Bias detection with the model difference procedure.

		Cond.	Biased items[a]						Unbiased items[b]					
			Conv.	Δ DIC			TP		Conv.	Δ DIC			FP	
				Q_{05}	Q_{50}	Q_{95}	Strict	Liberal		Q_{05}	Q_{50}	Q_{95}	Strict	Liberal
CONTINUOUS LATENT VIOLATOR														
$\rho(T,V) = 0.0$	No bias	1	–	–	–	–	–	–	0.93	−10	0	10	0.102	0.007
	Uniform	2	0.59	−100	−60	−40	1.000	1.000	0.86	−10	0	10	0.255	0.042
	Nonuniform	3	0.93	−110	−70	−30	1.000	0.989	0.92	−10	0	10	0.219	0.032
	Both	4	0.65	−188	−120	−82	1.000	1.000	0.88	−20	0	10	0.305	0.068
$\rho(T,V) = 0.5$	No bias	5	–	–	–	–	–	–	0.94	−10	0	10	0.115	0.012
	Uniform	6	0.93	−80	−40	−20	1.000	0.978	0.93	−10	0	10	0.263	0.045
	Nonuniform	7	0.96	−140	−80	−40	1.000	1.000	0.95	−10	0	10	0.213	0.023
	Both	8	0.92	−190	−120	−70	1.000	1.000	0.94	−20	0	10	0.383	0.097
CONTINUOUS OBSERVED VIOLATOR														
$\rho(T,V) = 0.0$	No bias	9	–	–	–	–	–	–	0.99	−4	1	5	0.232	0.002
	Uniform	10	0.91	−100	−67	−47	1.000	1.000	0.99	−9	0	5	0.485	0.028
	Nonuniform	11	0.92	−107	−67	−40	1.000	1.000	0.99	−9	0	4	0.451	0.032
	Both	12	0.92	−173	−137	−92	1.000	1.000	0.99	−12	−1	4	0.578	0.073
$\rho(T,V) = 0.5$	No bias	13	–	–	–	–	–	–	0.77	−5	2	5	0.274	0.009
	Uniform	14	0.52	−90	−50	−27	1.000	1.000	0.61	−13	−1	4	0.529	0.082
	Nonuniform	15	0.73	−121	−81	−47	1.000	1.000	0.76	−10	0	4	0.479	0.037
	Both	16	0.49	−167	−125	−94	1.000	1.000	0.64	−16	−3	4	0.672	0.172
DICHOTOMIZED OBSERVED VIOLATOR														
$\rho(T,V) = 0.0$	No bias	17	–	–	–	–	–	–	1.00	−4	1	5	0.239	0.008
	Uniform	18	0.99	−66	−42	−23	1.000	1.000	1.00	−9	0	4	0.390	0.032
	Nonuniform	19	1.00	−71	−43	−20	1.000	0.980	0.99	−7	0	4	0.401	0.028
	Both	20	0.98	−115	−84	−51	1.000	1.000	0.99	−10	0	4	0.462	0.044
$\rho(T,V) = 0.5$	No bias	21	–	–	–	–	–	–	0.94	−5	1	5	0.266	0.005
	Uniform	22	0.70	−53	−25	−10	1.000	0.943	0.86	−8	0	4	0.456	0.019
	Nonuniform	23	0.93	−79	−47	−21	1.000	1.000	0.93	−8	1	4	0.392	0.017
	Both	24	0.66	−116	−81	−55	1.000	1.000	0.83	−13	−1	4	0.550	0.099

Cond., Condition; Conv., proportion of converged solutions; Δ DIC denotes the difference in DIC between the reference model and the competing model; [a] Quantile DIC difference values (i.e., 05, 50, 95), and proportions of true positives (TP) are calculated over the converged solutions [of 1 (biased item) × 100 (replicates) = 100 solutions]; [b] Quantile DIC values (i.e., 05, 50, 95), and proportions of false positives (FP) are calculated over the converged solutions, which are 6 (non-biased items) × 100 (replicates) = 600 solutions in Conditions 1, 5, 9, 13, 17, and 21, and 5 (non-biased items) × 100 (replicates) = 500 solutions in all other conditions.

As can be seen in **Table 4**, the proportions of true positives (i.e., indicating bias when it is present) are very high in all conditions; both using the strict criterion (all 1.000) as using the liberal criterion (ranging from 0.943 to 1.000).

The quantiles of the DIC difference values give an indication of the power to identify items with bias. These DIC difference values are highly negative in conditions with both uniform and nonuniform bias, for all conditions with a continuous violator. In conditions with a dichotomous violator, we observe smaller negative DIC difference values, suggesting a lower power.

In all conditions with a continuous violator, the proportions of false positives with the strict cut-off value are high (ranging from 0.102 to 0.672), and with the liberal cut-off value reasonably low (maximally 0.172). For the dichotomized violator a similar pattern is observed. When applying the model difference procedure, the liberal cut-off value appears to perform better than the strict cut-off value, in terms of a proper balance between true positives and false negatives.

3.1.3. Sensitivity analyses

To assess the sensitivity to the choice of the priors for those parameters that express uniform and nonuniform bias, we reanalyzed the simulated data sets in the "most difficult" conditions: with both uniform and nonuniform bias and a correlated trait and violator (i.e., Conditions 8, 16, and 24) using clearly incorrect priors. That is, for the parameters expressing the bias, we use priors that reflect an absence of bias (i.e., a normal distribution with a mean of zero, for b and c). **Table 5** shows the measures of accuracy and efficiency of the estimated parameters, in a similar way as reported in **Table 2**. Comparing the results of **Tables 2**, **5** shows that, in case of a continuous violator, the estimation bias is still remarkably low when faced with clearly incorrect priors (all absolute values lower than 0.016). Also the efficiency and MCM standard errors and the convergence rates of these two conditions are comparable to those in **Table 2**. Also in case of a dichotomized violator, the estimation bias (with values −0.130 and 0.016) is very similar to the corresponding values in **Table 2**.

Table 5 | Estimation bias of the sensitivity analysis.

Violator	Cond.	Conv.	m($\overline{\theta_b}$) − b	sd($\overline{\theta_b}$)	m(se(θ_b))	m($\overline{\theta_c}$) − c	sd($\overline{\theta_c}$)	m(se(θ_c))
Continuous latent	8	0.95	−0.002	0.081	0.001	0.016	0.070	0.001
Continuous observed	16	0.48	0.014	0.058	0.001	0.006	0.058	0.001
Dichtomized observed	24	0.75	−0.130	0.056	0.001	0.016	0.061	0.001

Cond., Condition; Conv., proportion of converged solutions; All summary measures are calculated over the converged solutions only and are using the same notation as in ***Table 2****.*

Proportions of true and false positives hardly change when using the clearly incorrect prior, as has been verified (but not reported). We conclude that inadequate priors hardly influence the parameter estimates in all conditions, at least with the iteration length used in this simulation study.

4. DISCUSSION

In this article, we consider a Bayesian RFA approach for the detection of uniform and nonuniform bias. Results of a simulation study show that the parameter estimates of this proposed Bayesian RFA are reasonably accurate and efficient. With a dichotomized observed violator we find less accurate results, which is due to a loss of information and a reduction of the effect size. Our results thus support the validity of the well-know criticism on the median split (see, e.g., MacCallum et al., 2002). This suggests that the use of MGFA in cases with a continuous observed violator, with its associated necessity to dichotomize, should be discouraged. We used informative priors to obtain accurate and efficient results for the parameter estimates. Our sensitivity analysis shows that clearly inaccurate priors for the parameters expressing the bias also yield accurate and efficient estimates. This result might be different when working with a smaller sample than the $n = 500$ used in this paper. The smaller the sample size, the larger the influence of the prior distribution. In practice, to obtain priors we have to utilize prior information from different sources available (e.g., knowledge of experts or analyses of similar data), or perform an auxiliary estimation on a part of the data.

Results show that the Bayesian RFA is hindered by convergence problems, particularly in conditions with uniform bias. We used the Raftery and Lewis diagnostic to determine the number of iterations, but noticed in small experiments that doubling the number of iterations still decreased the number of convergence problems, according to the Gelman and Rubin diagnostic and Geweke diagnostics, substantially. For example, for condition 16 in **Table 2**, doubling the number of iterations increased the convergence rate from 52 to 69%. Thus there are indications that several of the convergence and estimation problems encountered in this simulation study, can be overcome in an empirical context through solutions such as choosing more chains, performing more iterations, and changing the initial values of the chains. Studying convergence properties for a variety of settings, including a variety of sample sizes, would be an interesting topic for future research.

The bias detection rates of both the single run procedure and the model difference procedure, calculated with either a strict or a liberal cut-off value, are very good. In both bias detection procedures, the distribution of the DIC difference values in the various conditions shows that the power to detect bias is the highest in conditions with a continuous observed violator. In conditions with a dichotomous observed violator there is a reduction of power, indicated by lower DIC difference values. In general, nonuniform bias is detected about as well as uniform bias is. However, if we focus on the DIC difference values, conditions with a independent trait and violator and nonuniform bias have smaller DIC difference values than conditions with uniform bias. In conditions with a dependent trait and violator, it is the other way round. This might be related to the fact that both the dependency between the trait and the violator and the bias are positive which may amplify each other.

Overall, the false positive rates are too large with a strict DIC cut-off value. Given the fact that a liberal cut-off value yields satisfactory bias detection results, we recommend a liberal DIC cut-off value (see also Lee, 2007). The false positive rates of the liberal DIC cut-off value are acceptable in all conditions and clearly lower in the model difference procedure. This might be due to a more precise estimation of the DIC difference procedure, as the model difference procedure directly compares a model with and without parameters to assess bias.

As an alternative to a liberal cut-off value, it might be helpful to detect bias in an iterative procedure. In this iterative procedure, the item associated with the largest DIC difference value is considered biased. In a second run, this bias is taken into account by allowing parameters that express bias in the model, and the bias test is conducted on the remaining items. As none of the remaining items is considered biased or half of the items are detected as biased, the iterative procedure stops (see Barendse et al., 2012, for an implementation in the frequentist framework).

Overall, for bias detection with the Bayesian RFA, both procedures with a liberal cut-off value are successful under the conditions studied. The model difference procedure appears to be more powerful in detecting bias and is therefore preferable over the single run procedure. The results presented indicate that the Bayesian RFA method is promising to assess measurement bias. It can be used to assess measurement bias with respect to multiple violators simultaneously, and of any kind (i.e., continuous or discrete, observed or latent).

For further research on the Bayesian RFA, it is useful to investigate model performance under other conditions, including larger numbers of observed items and varying the size of the bias. The size of the bias can be varied both in terms of severity and number of biased items. Additionally, more complicated models, with more than one item with bias, could be investigated. Further, extending the model with a latent categorical violator might be a useful extension. It may also be useful to consider alternative, promising, criteria for bias detection, such as Bayes factors and path sampling which both can deal with non-linearity (Lee,

2007). Finally, it would be highly interesting to see whether to theoretical advantages of Bayesian RFA are of use in empirical practice, by applying the methodology of this paper to empirical data.

AUTHOR CONTRIBUTIONS

All authors meet the criteria for authorship. All authors contributed substantially to the conception and design of the work, and drafting and finalizing the paper. M. T. Barendse, C. J. Albers, and M. E. Timmerman designed the simulation study, and M. T. Barendse and C. J. Albers programmed the simulation study.

ACKNOWLEDGMENT

This publication is supported by MAGW open competition grant 400-09-084 from the Netherlands Organization for Scientific Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.01087/abstract

REFERENCES

Barendse, M. T., Oort, F. J., and Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias; a simulation study. *AStA Adv. Stat. Anal.* 94, 117–127. doi: 10.1007/s10182-010-0126-1

Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., and Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Struct. Equ. Modeling Multidiscip. J.* 19, 561–579. doi: 10.1080/10705511.2012.713261

Bohrnstedt, G. W., and Goldberger, A. S. (1969). On the exact covariance of products of random variables. *J. Am. Stat. Assoc.* 64, 1439–1442. doi: 10.1080/01621459.1969.10501069

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2013). *Bayesian Data Analysis. 3rd Edn., Texts in Statistical Science Series*. London, UK: Chapman and Hall/CRC.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Geweke, J. (1992). "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in *Bayesian Statistics*, eds J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (New York, NY: University Press), 169–193.

Klein, A. G., and Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the lms method. *Psychometrika* 65, 457–474. doi: 10.1007/BF02296338

Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian Approach*, Vol. 711. Chichester: John Wiley and Sons. doi: 10.1002/9780470024737

Lee, S.-Y., and Song, X.-Y. (2012). *Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences*. Chichester: John Wiley and Sons.

Lunn, D., Spiegelhalter, D., Thomas, A., and Best, N. (2009). The BUGS project: evolution, critique and future directions. *Stat. Med.* 28, 3049–3067. doi: 10.1002/sim.3680

MacCallum, R. C., Zhang, S., Preacher, K. J., and Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychol. Methods* 7, 19–40. doi: 10.1037/1082-989X.7.1.19

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

MRC Biostatistics Unit. (2006). *Dic: Deviance Information Criterion*. Available online at: http://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-dic/ (accessed July 30, 2014).

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika* 54, 557–585. doi: 10.1007/BF02296397

Muthén, B. O., and Asparouhov, T. (2003). Modeling interactions between latent and observed continuous variables using maximum-likelihood estimation in Mplus. *Mplus Web Notes* 6, 1–9. Available online at: http://statmodel2.com/download/webnotes/webnote6.pdf

Muthén, B. O., and Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* 17, 313. doi: 10.1037/a0026802

Muthén, B. O., and Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus Web Notes* 17, 1–48. Available online at: https://www.statmodel.com/examples/webnotes/webnote17.pdf

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika* 6, 150–166.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Struct. Equ. Modeling Multidiscip. J.* 5, 107–124. doi: 10.1080/10705519809540095

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* 6, 7–11. Available online at: http://cran.r-project.org/doc/Rnews/Rnews_2006-1.pdf

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raftery, A. E., and Lewis, S. M. (1992). [practical markov chain monte carlo]: comment: one long run with diagnostics: implementation strategies for markov chain monte carlo. *Stat. Sci.* 7, 493–497. doi: 10.1214/ss/1177011143

Raftery, A. E., and Lewis, S. M. (1995). "The number of iterations, convergence diagnostics and generic metropolis algorithms," in *Practical Markov Chain Monte Carlo*, eds W. R. Gilks and D. J. Spiegelhalter (London, UK: Chapman and Hall), 115–130.

Roberts, G. O. (1996). "Markov chain concepts related to sampling algorithms," in *Markov chain Monte Carlo in practice*, eds W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (London, UK: Springer), 45–57. doi: 10.1007/978-1-4899-4485-6-3

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Tanner, M. A., and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* 82, 528–540. doi: 10.1080/01621459.1987.10478458

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 February 2014; accepted: 09 September 2014; published online: 29 September 2014.

*Citation: Barendse MT, Albers CJ, Oort FJ and Timmerman ME (2014) Measurement bias detection through Bayesian factor analysis. Front. Psychol. **5**:1087. doi: 10.3389/fpsyg.2014.01087*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in Psychology

METHODS
published: 27 March 2015
doi: 10.3389/fpsyg.2015.00289

A new visualization and conceptualization of categorical longitudinal development: measurement invariance and change

Jan Boom *

Developmental Psychology, Utrecht University, Utrecht, Netherlands

The Overlapping Waves Model (OWM) is a metaphor introduced by Siegler (1996) to illustrate a typical sequence of increasing and decreasing use of strategies during development. Going beyond metaphor, a new model synthesized from Latent Growth Modeling (LGM) and Item Response Theory (IRT) will be presented to analyze such categorical longitudinal data. Use of strategies can be scored as a variable with only a few ordinal categories. IRT provides the means to relate the usage of strategies to position on an underlying developmental dimension. LGM allows to model movement of individuals along this dimension, acknowledging individual differences both in starting point and in speed of progress. Measuring and modeling such strategy development requires that at each time point the same categories are used, in the sense that item difficulties must remain invariant over time. Whether, discrimination can be relaxed is still an issue. The problem that had to be solved was disentangling the between-person-individual differences from real intra-individual developmental differences. Figures with polytomous or multi-category Item Characteristic Curves (ICC's) resemble the OWM in many respects. However, such figures are usually taken to represent inter-individual differences, whereas the OWM usually represents development (so intra-individual differences), and we cannot have both at the same time. The solution came from creating a framework with ability differences on one axis and the effect of time on another axis, resulting in a 3-D model. These (orthogonal) dimensions make it possible to adequately conceptualize measurement invariance in this complex context. As the result is difficult to conceptualize without extensive visualization, special 3-D figures will be used to illustrate and a dynamic (rotatable and scalable) version will be made available as Computable Document Format object (Mathematica). The model was successfully applied in several microgenetic studies.

Keywords: measurement invariance, strategy development, overlapping waves, latent growth modeling, item response theory

OPEN ACCESS

Edited by:
Peter Schmidt,
University of Giessen, Germany

Reviewed by:
Bobby Naemi,
Educational Testing Service, USA
Daniel Seddig,
University of Zürich, Switzerland

***Correspondence:**
Jan Boom,
Developmental Psychology, Utrecht University, Heidelberglaan 1, 3584 CS Utrecht, Netherlands
j.boom@uu.nl

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *09 July 2014*
Accepted: *27 February 2015*
Published: *27 March 2015*

Citation:
Boom J (2015) A new visualization and conceptualization of categorical longitudinal development: measurement invariance and change. Front. Psychol. 6:289. doi: 10.3389/fpsyg.2015.00289

Introduction

Measurement invariance (MI) is mostly considered in the context of differences between subpopulations (inter-individually), however, measurement invariance is also important in a longitudinal

context. It might be unfair if an instrument does not measure the same constructs for all subpopulations. However, in a longitudinal study, if we compare the same sample (ignoring attrition for the moment) with itself on different occasions, the issue is perhaps not fairness but whether the study is effective in being able to identify progress.

The case I want to focus on concerns strategy development. With development individuals might move from using one strategy to another. Such strategies might be qualitatively different and hierarchically ordered. Let's suppose they are and that the following assumptions about the underlying structure of the data hold to a sufficient degree.

The first assumption is that the strategies can be ordered in terms of advancedness. Higher numbered strategies are better (although it might be difficult to define exactly in what sense) such that each next strategy may become attractive once the presently used one is sufficiently mastered. The second assumption is that participants use only one strategy at a time. This assumption does not necessarily contradict what has become received view: that there is—and has to be—a broad variation in strategies from which—like in evolution—the best strategies are chosen (Siegler, 1996). However, even such a view implies that there are different strategies. Thus, while participants may have several strategies at their disposal, the assumption is that one problem is solved, in the end, by using only one strategy. The next moment, or the next problem, may involve the use of other strategies, so in that sense "use" still might be a mixture. The third assumption is closely related to the previous two and holds that there is a single underlying dimension which represents advancedness of these strategies. This implies that the strategies will map on an ability scale with strategies as markers along this scale and defining the scale. The intervals between strategies along the scale need not be regular but the scale is assumed to be one-dimensional only. Using a strategy can now be scored as an ordinal variable with few categories and longitudinal development as a vector of such scores per participant. These assumptions are compatible with many classical developmental theories: developmental scales, stage theories, skills theory, and hierarchical complexity theory.

I have proposed a formal statistical model to analyze such data by connecting group level trends to an underlying developmental dimension valid on the individual level. Let me explain first why this is has been a problem so far in developmental psychology before returning to MI and details.

The Overlapping Waves Model (OWM) was introduced by Siegler (1996) as a metaphor to illustrate the typical pattern for many cognitive tasks of a sequence of increasing and decreasing use of strategies, or rules as he called them, during *individual* development (**Figure 1**). Such a pattern might apply, for example, to children learning to multiply numbers below 10: Strategy 1 might refer to incorrect approaches such as guessing; Strategy 2 might refer to finger counting; A more advanced strategy is repeated addition; The most advanced strategy in this example is retrieval from memory. Compare this model to results from a famous longitudinal study by Colby et al. (1983) on stage-wise moral development, in which they reported increasing and decreasing use of five stages of moral development for the group level (**Figure 2**).

A fundamental problem, that has plagued developmental theorizing since long, is that it is difficult to infer the shape of development from group results. **Figure 2** refers to actual empirical data, but trends are only valid on the group level; whereas, **Figure 1** suggests being valid on the individual level, but does not directly reflect empirical data (it's just hypothetical). On the one hand, because the shape of non-linear trends need not be comparable between group and individual data, referring to group data as in **Figure 2** will not do as support for claims about typical *individual* trajectories. On the other hand, actual individual data (trajectories), as e.g., presented abundantly in Siegler (1996) and Colby et al. (1983), have not been used to formally confirm trends as in **Figure 1**, possibly because the overwhelming individual variation. Of course, there are likely to be constraints and relationships between the individual and the group level, but a formal model of the exact nature of these relationships was lacking so far.

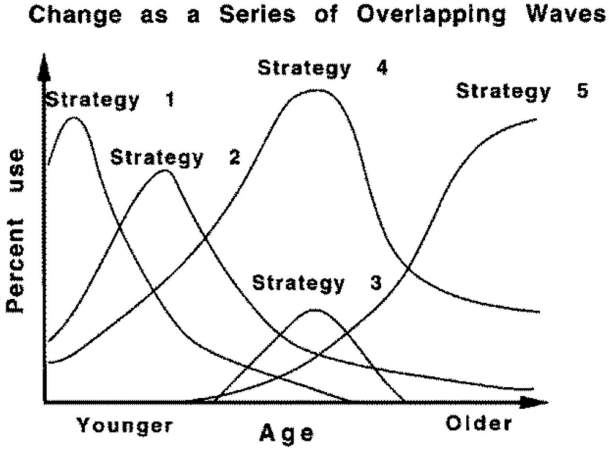


FIGURE 1 | Overlapping waves depiction of cognitive development. From "Emerging Minds," by Siegler (1996).

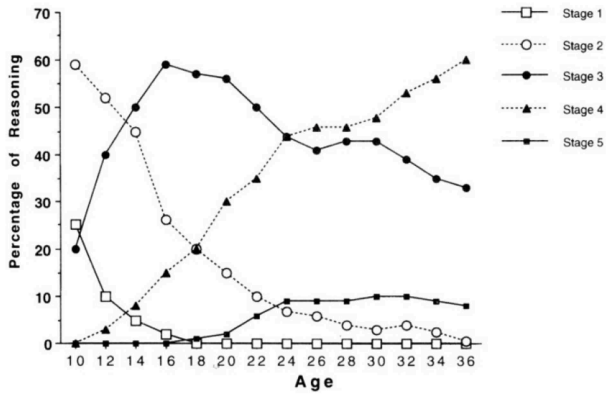


FIGURE 2 | Percentage use of stages of moral reasoning. Data from Colby et al. (1983).

Going beyond metaphor, I developed a new conceptualization synthesized from Latent Growth Modeling (LGM) and Item Response Theory (IRT) to understand such categorical longitudinal data. The model itself is not new, because Muthén explored such models extensively (Muthén, 1996; Muthén and Asparouhov, 2002, 2013), but the application to strategy development is entirely new. IRT provides the means to relate the likelihood of use of particular strategies to a position on an underlying developmental dimension. LGM allows modeling movement of individuals along this dimension, acknowledging individual differences both in starting point and in speed of progress (and more).

Measurement Invariance must hold in such models in the sense that strategies themselves do not change with time or age. What is supposed to change is the use, or the propensity to use them. MI might be violated if a new factor has become influential over time, the result would be that we cannot find progression developmentally. Whether discrimination can be relaxed is an issue to be discussed below. I will first introduce IRT and LGM.

Item Response Theory Modeling

With the assumptions in place Item Response Theory (IRT) provides the means to relate the use of strategies to an underlying ability.

Suppose responses of participants are scored as representing the use of one of a few strategies, suppose furthermore -following IRT modeling- that the likelihood of using one of the strategies depends on a single latent variable by a mathematical function known as the item response function. The Partial Credit Model is particularly useful to model responses that are ordered as a series of steps that must be mastered in sequence (Millsap, 2010). The latent variable normally represents inter-individual differences in ability, which in this case translates to being more or less advanced in terms of strategy use as is illustrated in **Figure 3**. Figures are based on a micro-genetic study (Van der Ven et al., 2012) and just used to illustrate here. The multiplication strategies examples given earlier are from this study.

The X-axis of **Figure 3** can be thought of as representing strategy ability differences between participants (as person characteristic). In that case a position more to the right represents a higher ability, more to the left represents less ability. On the Y-axis is the likelihood of using the particular strategy. For example, likelihood of using strategy-5 is higher for persons with high strategy ability, while strategy-1-use diminishes rapidly with increasing ability.

Alternatively, the X-axis can be used to represent characteristics of the strategy; e.g., the peak of each of the middle curves gives the most typical ability value for that strategy, but it is also clear there is considerable overlap between strategies.

The result is a latent strategy ability scale that can be used to represent inter-individual ability differences and strategy advancedness. The position along this strategy scale is nonlinearly and probabilistically related to the use of the various strategies. The attractiveness of the transformation of the categorical scores to this unbounded continuous interval scale is that it opens up the possibility to use all kinds of regression techniques.

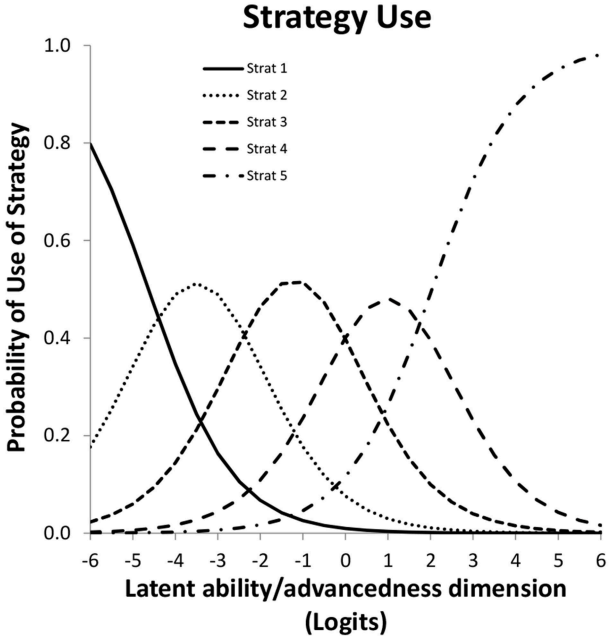


FIGURE 3 | Item characteristic curves for a five-category polytomous item.

However, one of the key ideas of this paper is that the X-axis can also be used to represent intra-individual development, as in Siegler's Overlapping Waves Model. In other words: also development over time can be conceptualized and visualized as a shift to right in **Figure 3**. The result in terms of expected strategy use can be quite complex to describe because it depends on the starting point and the growth rate of the particular subject. Only the first and the last strategy have a consistent change pattern, the use of all other strategies goes up and down. The profile of shapes depends on properties of the item and may be different for different items. For a Partial Credit model version the basic shape (steepness of the curves) is fixed. Although scaling of X-axis is arbitrary, location (to the right or left) can vary between items, and applies to the whole set of curves for an item. Height of the curves, or area's beneath it, which can also be expressed as distance between crossings of curves, may be different within or between items.

Basic principles of relevant IRT modeling, and some alternative models, are reviewed by de Ayala (2009) and Embretson and Reise (2000), more details on polytomous item response models can be found in Ostini and Nering (2006), and more on categorical data-analysis in general in Agresti (2002).

Latent Variable Growth Curve Modeling

Whereas, IRT provided the means to relate the use of strategies to an underlying dimension, development of individuals along this underlying dimension can be modeled by means of Latent variable Growth curve Modeling (LGM). LGM is a powerful and flexible technique, which can be used to model longitudinal development (Bollen and Curran, 2006; Duncan et al., 2006).

A linear LGM, for example, presupposes a steady increase or decrease in the target variable over a small number of equally spaced measurement occasions for each person. The increase is assumed to be linear, with an intercept and a slope parameter describing a trajectory for each individual. These intercepts and slopes are assumed to be different for each respondent and normally distributed in the population with unknown mean and variation. In the usual continuous case, the observed scores for an individual participant will depart from his or her best fitting straight line, and it is assumed that these residuals are normally distributed in the population with zero mean and certain variance for each measurement occasion; moreover, these residuals are not correlated over measurements occasions. However, in the present case, where we combine the LGM with an IRT model, the role of residuals can be treated in other ways too. Because this has consequences for the issue of MI I will return to the role of residuals in a moment.

Concluding: a linear LGM might be suitable to model increasing strategy development over measurement occasions, and, if the model holds, every participant's trajectory can be represented by a straight line, as will be shown shortly.

Three Dimensional Overlapping Waves Model

The Overlapping Waves model of **Figure 1** was originally presented to visualize development, whereas the polytomous item response model of **Figure 3** is more commonly used to depict individual differences. With a three-dimensional version of the Overlapping Waves model both uses can be combined.

In **Figure 4** the X-axis refers to individual differences, the Y-axis to time (measurement occasions = 8 weeks in our empirical example), and the Z-axis refers to probability of using one of five strategies. The floor is a two dimensional plane on which a growth curve model can be placed as illustrated in **Figure 5**.

In **Figure 5** estimated (idealized) individual trajectories of development in strategy use are shown on the floor plane for a subsample of 20 participants. For one individual, as illustration, the implied category boundaries are also depicted in the Z-plane. The degree of curvature is limited, which makes sense, because it represents only a modest increase on the strategy maturity scale as in **Figure 3**. **Figures 4**, **5** are combined in **Figure 6** to illustrate that each individual curve from **Figure 5** follows the surface of the waves in **Figure 4**.

The more growth, in **Figure 6**, the more the set of curves for a particular individual turns away from an orientation parallel to the week axis, and the more curvature (in the heavy black lines) will result. No growth, as e.g., for the case with the line most to the left on the floor of **Figure 5**, would result in straight heavy black lines for each of the 5 strategies when projected on the surfaces. But also a different starting point (different intercept in the growth model part) can lead to completely different curvatures: imagine the set of curves being shifted along the difficulty dimension.

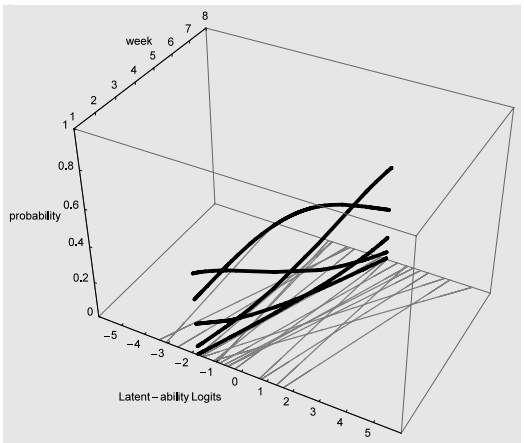


FIGURE 5 | LGM trajectories (thin gray lines) and snapshot of Item Characteristic Curves for one participant (heavy black lines).

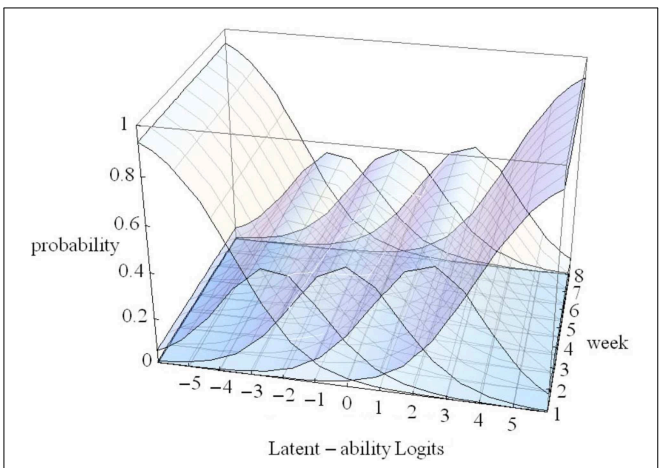


FIGURE 4 | 3D-Overlapping waves model (see text).

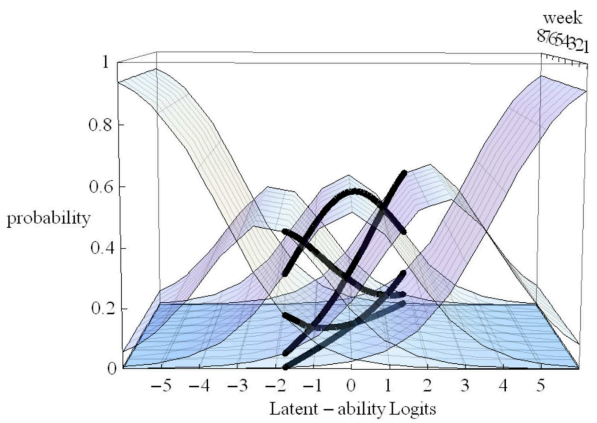


FIGURE 6 | 3D-Overlapping waves model with item characteristic curves for one participant as illustration.

Measurement Invariance Revisited

Since we are interested in development of using strategies over time some assumptions had to be made. In principle an assumption cannot be tested (not directly at least). Scoring observations or verbal material etc. needs to be done first and this often implies a considerable amount of preprocessing of the raw data. Categories have to be defined in advance, so that judges can apply these to the observations. The categories must be defined such that they are ordered in terms of difficulty. Age of respondents whose responses are to be assigned to the categories cannot and may not play a role at all in the definition of the categories. Nor should time or measurement occasion play a role in the definition of the categories. Therefore, if we relate these categories to an underlying ability this relation must remain strictly invariant over measurement occasions.

Specifying the relation between categories and underlying ability can be done in many ways, but always involves the difference between the latent ability score and thresholds. A threshold τ_j is the value on the scale where the likelihood for being assigned category changes from being greater for j to being greater for $j+1$ (so around $p = 0.5$) and represent what in IRT parlance would be the difficulty of the item.

For a weighted least-squares (WLS) estimator with Probit link and Theta parametrization the Item Characteristic Curves (ICC) is specified as follows: Let U_i be a categorical indicator for a latent ability factor f with categories $j = 0, 1, 2,..., J$-1, for item $i = 1, 2,... I$, and measurement occasion $t = 1, 2,... T$.

$$\begin{aligned} P_{itj}(f) &= P(U_{it} = j|f) \\ &= \Phi\left(\frac{\tau_{itj} - \lambda_{it}f}{\sqrt{\theta_{it}}}\right) - \Phi\left(\frac{\tau_{itj-1} - \lambda_{it}f}{\sqrt{\theta_{it}}}\right) \end{aligned} \quad (1)$$

Where Φ is the standard normal distribution function, τ_j is the threshold for category j, λ is the factor loading, θ is the residual variance. For the first category the second Φ term is zero, for the last category the first Φ term is 1. Note, however, that Mplus offers not only WLS estimators but also maximum likelihood (ML) estimators, not only Probit links but also Logit links, and also a Delta parametrization (see Muthén, 2010 for an overview), and for each case the ICC's are differently specified. **Figure 3** is an example of a ML Logit ICC, however, apart from scaling differences the Figure would be almost the same as the one obtained from formula 1.

Regarding difficulty; with more items and more measurements occasions the thresholds τ_{ij} (for each separate category) are to remain invariant over occasions, but may be different over items i. Steps may be defined as $s_j = \tau_{ij} - \tau_{ij-1}$ and restricted, or not, to be equal between items.

Regarding discrimination; dividing by the standard deviation of the residuals θ_{it} in formula 1 allows introducing differences in discriminations over items and occasions. With ML estimators this is difficult to achieve and discrimination differences are not implemented in Mplus for ML. Being able to allow differences (with WLS) in relative discrimination between items is however an attractive option that can improve fit. Whether, allowing the residual variance θ_{it} to be variant over measurements occasions, as is advocated by Muthén and Asparouhov (2002), is a good idea has to be seen. It will undoubtedly improve fit but interpretation might be more difficult. In **Figures 3**, **5** for some occasions the Figure will in that case be broader or slimmer which is detrimental to the intended general applicability of the model: As outlined above the intention is to be able to handle large individual differences in ability with this model.

Example: Stepwise Understanding Randomness

The general model has been applied to several data sets now (e.g., Van der Ven et al., 2012). The pilot study presented below is just meant as an illustration.

Children's understanding of randomness was studied using a microgenetic design in three age cohorts (Grade 1, 3, and 5) of different primary schools in rural parts of the Netherlands. During 5 weeks four probability-related questions about a marble tilt box (see Metz, 1998) were administered weekly to 75 children. A box 30 cm wide and 40 cm long, with edges 5 cm high was mounted on a support of 5 cm high, affixed to the bottom, such that it could be tilted from one side to the other such that marbles would roll from one side to the other. Initially all marbles were lying in row on the lower side: 5 white on the left and 5 green on the right. Questions were asked before, during, or tilting of the box: e.g., "How will the marbles end up on the other side?"; "Can you be sure"; "What happened?" (After trying it out themselves); "What if we did tilt the box a 100 times?"; "Can the original distribution of the marbles (5 and 5 neatly separated) ever occur again?" Answers given by each child were coded using 4 possible categories (based on Metz, 1998). Developmental progress was presumed to go from: (1) No understanding at all to (2) determinism which is denial of chance element (they have to go back to their places), to (3) unpredictability (you never know, they just roll), to (4) recognizing some degree of long term predictability (returning is possible but unlikely). Teaching about randomness is not a part of curriculum in the Netherlands. Participants were not given feedback by the test administrators, but were able to see, of course, the outcomes when the task was eventually played each week. This resulted in a 75 (participants) by 16 (variables = 4 question-sets by 4 weeks) raw data-matrix with codes one to four.

Results

Figure 7 shows that the four items (= topic/question-set) had different profiles. These shapes of the item profiles are fixed over the 4 weeks (to achieve measurement invariance). On the x-axis is the difficulty of the item (to use the IRT parlance) which in this case reflects whether the particular question tends to elicit more advanced or more simplistic responses. Note that children can be placed on the same x-axis scale: see **Figure 8**.

Figure 8 shows the expected changes in the use of the categories (levels of understanding), over weeks, for each item (topic/question-set), separately. The scale on the x-axis corresponds exactly to **Figure 7**. The actual abilities, in this case, cover only small part of the scale (the scale is centered around zero because the average is arbitrarily set to zero).

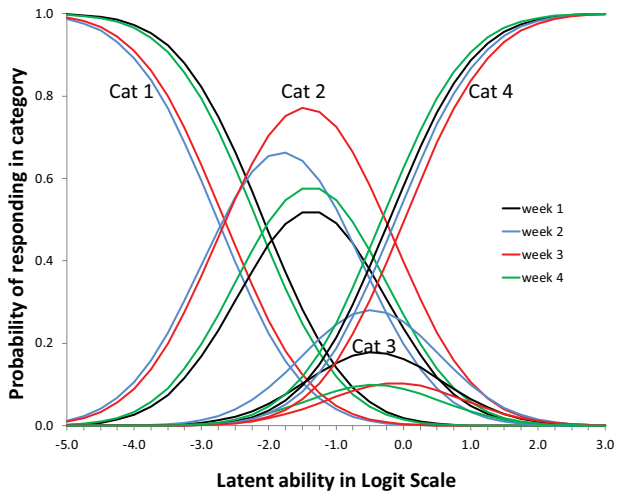


FIGURE 7 | Item profiles for four topic/question sets. On the x-axis is the latent ability/advancedness scale, as explained earlier.

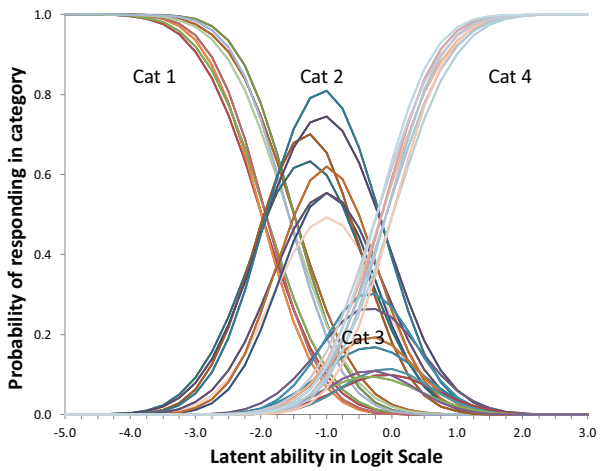


FIGURE 9 | Item profiles for four topic/question sets, while allowing for some measurement differences between weeks. On the x-axis is the latent ability/advancedness scale, as explained earlier.

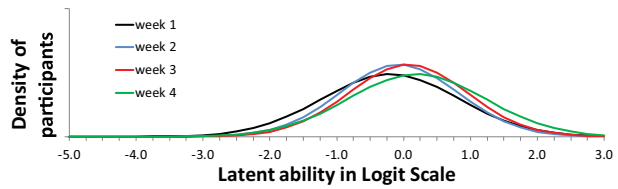


FIGURE 8 | Model Implied (expected) week trends of use of categories (levels of understanding).

Increase over weeks (the slope in the growth model part) was significant. Nevertheless, seen from a substantive viewpoint, the result are not all good. Category three has very low occurrence (never dominant) and might be better removed from the coding scheme. The range of respondent abilities is not corresponding to the item categories very well. No substantive conclusions from this illustration can be drawn. However, seen from a modeling and analysis viewpoint results are good and interesting. Mplus 7.2 was used to estimate and fit all models (Muthén and Muthén, 1998-2012). Fit for the WLS Probit version with DELTA parameterization was acceptable with an RMSEA of 0.075; CFI of 0.911; TLI of 0.930. More options for analysis were tried out (e.g., MLR-Logit, WLS-THETA) and all converged to the same kind of Figures (as in **Figures 7**, **8**). The analysis demonstrates that with a rather small sample already interesting results are possible and weaknesses in the data or coding scheme are revealed without fail.

Regarding measurement invariance it might be argued that there might be systematic differences over 4 weeks, e.g., due to slightly different testing conditions. Using the scaling option in the Mplus DELTA parameterization we allowed scaling differences between measurement points. The resulting scaling factors were 1 (as anchor) for set 1 and 1069, 0.890, and 1.093, respectively for sets 2–4. Fit for this WLS Probit version with DELTA parameterization was almost the same with an RMSEA of 0.074; CFI of 0.914; TLI of 0.932. The scaling option in the DELTA parameterization gave slightly better results than the option to relax the discrimination (requiring the THETA parameterization) as mentioned earlier, but has the same kind of effect: introducing mild between measurement differences by just stretching or contracting the scale a bit. The result is **Figure 9** with four times as many, and more cluttered, lines than **Figure 7**.

Discussion

A new 3D Overlapping Waves model is presented, based on a combination of Latent variable Growth curve Modeling (LGM) and Item Response Theory (IRT) modeling. The statistical principles used are long established and sound. It is a formal model for conceptualizing strategy development which throws new light on the issue of variability and measurement invariance in development. It is also an empirically testable model which might be helpful in longitudinal studies in which the responding changes fundamentally over development or experience.

All advantages of LGM apply. Predictors can be added to the LGM part, it is also possible to test nonlinear growth, or add more growers to the model. All advantages of IRT modeling also apply. The new part is: that what is normally the end result (estimated individual scores or group indicators thereof) now is -in an additional step- transformed in a set of thresholds and these can be visualized as a set of curves (with strong shape constraints). More hypotheses concerning the relationship between strategies can be tested by specifying equality constraints between thresholds. Also the relationship between items (e.g., more or less difficult ones) can be further investigated and tested. IRT analyses are often based on much larger datasets; large item banks, and focused on item selection for a test. The present application of IRT is different and more experience with dealing with complex models with relatively few participants is needed.

The raw data may appear incredibly complex and variable, but, as shown, it may still be the case that the data are generated by relatively simple linear growth for each person reflecting an

underlying dimension of development. Of course, in actual practice there will always be violations of the assumptions, for all kinds of reasons, but if there is a consistent pattern over individuals to a sufficient degree, such an underlying dimension is plausible. The analysis can be done with commercially available software and is not difficult to conduct, although it requires some conceptual work and spatial imagination. The model has important theoretical implications!

Regarding measurement invariance over weeks, the illustrative example showed that relaxing the strict measurement invariance, by allowing some overall scaling differences, did not lead to serious improvement in fit, but since all the threshold are different the Figure is more difficult to understand and also difficult to compare to a more restricted model. Although more studies are needed before final recommendations can be given, a more parsimonious model seems preferable.

References

Agresti, A. (2002). *Categorical Data Analysis, 2nd Edn*. Hoboken, NJ: Wiley.

Bollen, K. A., and Curran, P. J. (2006). *Latent Curve Models: A Structural Equation Perspective*. Hoboken, NJ: Wiley.

Colby, A., Kohlberg, L., Gibbs, J., Lieberman, M., Fischer, K., and Saltzstein, H. D. (1983). A longitudinal study of moral judgment. *Monogr. Soc. Res. Child Dev.* 48, 1–124. doi: 10.2307/1165935

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: The Guilford Press.

Duncan, T. E., Duncan, S. C., and Strycker, L. A. (2006). *An Introduction to Latent Variable Growth Curve Modeling, 2nd Edn*. Mahwah, NJ: Erlbaum.

Embretson, S., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum.

Metz, K. E. (1998). Emergent understanding and attribution of randomness: comparative analysis of the reasoning of primary grade children and undergraduates. *Cognit. Instr.* 16, 285–365. doi: 10.1207/s1532690xci1603_3

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: an introduction. *Child Dev. Perspect.* 4, 5–9. doi: 10.1111/j.1750-8606.2009.00109.x

Muthén, B. O. (1996). "Growth modeling with binary responses," in *Categorical Variables in Developmental Research*, eds A. von Eye and C. Clogg (San Diego, CA: Academic Press), 37–54. doi: 10.1016/b978-012724965-0/50005-5

Muthén, B. O. (2010). *IRT in Mplus*. Available online at: http://www.statmodel.com/download/MplusIRT2.pdf

Muthén, B. O., and Asparouhov, T. (2002). *Latent Variable Analysis with Categorical Outcomes: Multiple-Group and Growth Modeling in Mplus*. Mplus Web Notes: No. 4. Available online at: http://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, B. O., and Asparouhov, T. (2013). *Item Response Modeling in Mplus: A Multi-Dimensional, Multi-Level, and Multi-Timepoint Example*. Available online at: http://www.statmodel.com/download/IRT1Version2.pdf

Muthén, L. K., and Muthén, B. O., (1998-2012). *Mplus User's Guide, 7th Edn*. Los Angeles, CA: Muthén & Muthén.

Ostini, R., and Nering, M. L. (2006). *Polytomous Item Response Theory Models*. Thousand Oaks, CA: Sage.

Siegler, R. S. (1996). *Emerging Minds: The Process of Change in Children's Thinking*. New York, NY: Oxford University Press.

Van der Ven, S. H. G., Boom, J., Kroesbergen, E. H., and Leseman, P. M. (2012). Microgenetic patterns of children's multiplication learning: confirming the overlapping waves model by latent growth modeling. *J. Exp. Child Psychol.* 113, 1–19. doi: 10.1016/j.jecp.2012.02.001

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 08 September 2014
doi: 10.3389/fpsyg.2014.00984

Measuring hedonia and eudaimonia as motives for activities: cross-national investigation through traditional and Bayesian structural equation modeling

Aleksandra Bujacz[1,2]*, Joar Vittersø[2], Veronika Huta[3] and Lukasz D. Kaczmarek[1]

[1] *Faculty of Social Sciences, Institute of Psychology, Adam Mickiewicz University, Poznañ, Poland*
[2] *Department of Psychology, University of Tromsø, Tromsø, Norway*
[3] *School of Psychology, University of Ottawa, Ottawa, ON, Canada*

Edited by:
Rens Van De Schoot, Utrecht University, Netherlands

Reviewed by:
Jan Ciecіuch, Cardinal Stefan Wyszynski University in Warsaw, Poland
Sebastiaan Rothmann, North-West University, South Africa

***Correspondence:**
Aleksandra Bujacz, Faculty of Social Sciences, Institute of Psychology, Adam Mickiewicz University, ul. Szamarzewskiego 89AB, Poznań 60-568, Poland
e-mail: aleksandra.bujacz@amu.edu.pl

Two major goals of this paper were, first to examine the cross-cultural consistency of the factor structure of the Hedonic and Eudaimonic Motives for Activities (HEMA) scale, and second to illustrate the advantages of using Bayesian estimation for such an examination. Bayesian estimation allows for more flexibility in model specification by making it possible to replace exact zero constraints (e.g., no cross-loadings) with approximate zero constraints (e.g., small cross-loadings). The stability of the constructs measured by the HEMA scale was tested across two national samples (Polish and North American) using both traditional and Bayesian estimation. First, a three-factor model (with hedonic pleasure, hedonic comfort and eudaimonic factors) was confirmed in both samples. Second, a model representing the metric invariance was tested. A traditional approach with maximum likelihood estimation reported a misfit of the model, leading to the acceptance of only a partial metric invariance structure. Bayesian estimation—that allowed for small and sample specific cross-loadings—endorsed the metric invariance model. The scalar invariance was not supported, therefore the comparison between latent factor means was not possible. Both traditional and Bayesian procedures revealed a similar latent factor correlation pattern within each of the national groups. The results suggest that the connection between hedonic and eudaimonic motives depends on which of the two hedonic dimensions is considered. In both groups the association between the eudaimonic factor and the hedonic comfort factor was weaker than the correlation between the hedonic pleasure factor and the eudaimonic factor. In summary, this paper explained the cross-national stability of the three-factor structure of the HEMA scale. In addition, it showed that the Bayesian approach is more informative than the traditional one, because it allows for more flexibility in model specification.

Keywords: hedonia, eudaimonia, well-being, Bayesian structural equation modeling, measurement invariance

INTRODUCTION

The distinctions between hedonic and eudaimonic notions of happiness has attracted a rapidly expanding group of well-being researchers (e.g., Keyes et al., 2002; Kopperud and Vittersø, 2008; Berridge and Kringelbach, 2011; Henderson et al., 2013; Huta, 2013; Huta and Waterman, 2013; Oishi et al., 2013; Ryan et al., 2013; Bauer et al., 2014; Proctor et al., 2014; Uchida et al., 2014). Both concepts are derived from ancient philosophy and they came to prominence within positive psychology as a result of research that conceptualized well-being in different ways. Proponents of a strict hedonic approach argue that a good life is properly accounted for by the presence of pleasure and the absence of pain (Kahneman, 1999; Tännsjö, 2007), whereas a broader approach includes positive attitudes and life satisfaction within the idea of hedonia (e.g., Diener, 1984; Feldman, 2004; Diener et al., 2009). By contrast, proponents of eudaimonic approaches believe that there is more to a good life than pleasant feelings and favorable attitudes (Tatarkiewicz, 1976; Ryff, 1989; Waterman, 1993; Ryan and Deci, 2001; Deci and Ryan, 2008; Keyes and Annas, 2009; Vittersø, 2013). In the current study, we enter this debate by looking into an established self-report scale that measures both hedonic and eudaimonic conceptions of well-being—the Hedonic and Eudaimonic Motives for Activities scale (HEMA; Huta and Ryan, 2010). Our two major aims are, first to examine the cross-cultural consistency of the factor structure of the HEMA, and second to illustrate the advantages of using Bayesian estimation for such an examination.

Existing attempts to quantify the associations between elements of hedonic and eudaimonic well-being show mixed results. In a recent review, Huta and Waterman (2013) attributed some of these inconsistencies to conceptual disagreements. Four categories of conceptualizations were identified, as Huta and Waterman observed that well-being has interchangeably been studied as orientations, as behaviors, as experiences and as

functioning. Another distinction in the well-being literature relates to the level of analyses: both trait level measures and state level measures were frequently observed. In order to avoid some of these confusions Huta and Waterman argued that it is important to clearly specify which category of analysis and level of measurement is taken into consideration. When focusing specifically on the distinction between eudaimonia and hedonia (as in a factor analysis) and on the stability of the correlation between them (as in a multigroup investigation) the confounding effect of different conceptualizations should be avoided. Therefore, the current paper operationalizes eudaimonia and hedonia as orientations, measured at the trait level. The HEMA scale fulfills both these criteria and was thus elected as the measurement instrument of our study.

The hedonic subscale of the HEMA addresses the two concepts that appear in most conceptions of hedonia: pleasure and absence of pain (e.g., Kahneman, 2000). The absence of pain is assessed in approach terms (as "seeking relaxation" and "seeking to take it easy") rather than in avoidance terms to minimize the confounding role played by the differential effects of approach and avoidance motivation (for a review see Elliot, 2008). This corresponds with the argument that pleasure—a proactive search for positive experiences—should be distinguished from comfort defined as a state of biological indifference (Scitovsky, 1976; Cabanac, 2010). For this reason the hedonic scale could actually be divided into the two dimensions of seeking relaxation and seeking pleasure (Asano et al., 2014).

The eudaimonic subscale of the HEMA covers three concepts which focus on personal qualities: authenticity (self-knowledge, autonomy, and integrity), excellence (virtue, performing to a high standard), and growth (learning, actualizing one's unique potentials, and maturing as a person). This operationalization is not exhaustive, as other concepts are often emphasized in definitions of eudaimonia (e.g., meaning, engagement). Nevertheless, the advantage of the HEMA scale is that it allows for a simultaneous assessment of both hedonic and eudaimonic orientations (measured together as motives), and therefore provides an opportunity to study the connection between them.

The relationship between factors of stable hedonic and eudaimonic orientations to well-being is hardly ever studied across different samples or national groups. Therefore, we don't know whether a particular figure representing the correlation between the factors is universal or specific for a national sample or a language version. To confirm a stability of connections between well-being constructs multigroup studies are needed. Yet for such designs the issue of measurement invariance (MI) becomes a crucial concern (e.g., Brown, 2006). This means that when a measurement tool is used across groups, its internal structure should follow at least two requirements: (1) the same number of factors should occur in all groups (configural invariance), and (2) the similar pattern of factor loadings should be observed across groups (metric invariance). If a model that imposes both of those requirements fits the data well, structural parameters—such as factor correlations—can be legitimately examined and compared across groups (e.g., Meredith and Teresi, 2006; Raykov et al., 2012). Additionally, when a comparison between latent means is of interest, the similar pattern of item intercepts should be established (scalar invariance).

In sum, the aim of this paper is to provide a systematic investigation of the correlational nature of the HEMA scale in two different nations. A confirmatory and multigroup factor analytic design was chosen for this purpose.

THE APPLICATION OF BAYESIAN ESTIMATION

With cross-national data from the HEMA scale, the analysis presented in this paper utilizes and compares two different estimation methods: (1) a traditional frequentist approach with maximum likelihood (ML) estimation and (2) a relatively new technique based on Bayesian structural equation models (BSEM) (Muthén and Asparouhov, 2012). This double analyses strategy was chosen in order to compare the results of those two methods and thereby provide an example that will reveal the possible advantages offered by Bayesian estimation. Since computational power nowadays supports the use of the Bayesian approach, it has been widely recommended due to the fundamental advantages of this method. Several introductory discussions of Bayesian estimation and inference exist (e.g., van de Schoot et al., 2013a; Zyphur and Oswald, 2013). The possible advantages of using BSEM can be found in all three steps of the analysis reported here.

First, BSEM allows the replacing of exact zero constraints with approximate zero constraints for different parameters of a model such as cross-loadings or residual covariances (Muthén and Asparouhov, 2012, 2013). This is possible due to the specific assumptions underlying Bayesian estimation. Bayesians treat parameters as variables characterized by a distribution, in contrast to the frequentist approach in which samples have distributions while parameters are fixed in the population (Zyphur and Oswald, 2013). Moreover, in Bayesian analysis a distribution for each of the parameters can be restricted by specifying priors, which are usually based on previous knowledge. For example, in confirmatory factor analyses (CFA) it is often assumed that each of the items will load on one factor only, hence the errant loadings (cross-loadings) are fixed at zero. However, the precise zero constraint has been criticized as unreasonable and unnecessary, because researchers usually want those errant-loadings to be very small (e.g., Golay et al., 2013). In most cases it may be enough to state that cross-loadings do not exceed a particular value, for example 0.3 (Brannick, 1995). Bayesian estimation allows us to place such a constraint by specifying a prior distribution for a cross-loading, in order to have little variance around the mean set to zero (i.e., an informative prior, e.g., van de Schoot et al., 2013b). Thanks to this option, the model fit will not suffer from an unreasonable assumption that does not reflect the true intention of the researcher.

Secondly, the same advantage of an approximate equality, rather than a precise one, can be employed for the MI analyses (Muthén and Asparouhov, 2013; Cieciuch et al., 2014). Traditional MI strategy places strong constraints on the parameters of a scale by forcing them to be identical across groups. Such an approach often leads to the conclusion that a scale is not invariant across groups, with little information about how big the differences are. Previous attempts to deal with this problem by establishing partial MI models remain controversial (e.g., Byrne

et al., 1989; Millsap and Kwok, 2004; Schmitt and Kuljanin, 2008). In this study, as in many others, the goal is to show that a scale performs in a very similar way across national groups. Yet, it is not expected that any particular item will behave differently across the groups (this could be solved by a partial MI model). Instead, we assume that all the items in the scale may vary across the nations and the size of these differences is of interest. BSEM allows the estimation of their magnitude by specifying limits for their distribution (which are set up by the informative priors). Thus, we decided to employ an approximate MI approach based on BSEM, and assumed that small deviations (i.e., statistically insignificant) would not jeopardize the comparison between factor covariances.

Thirdly, Bayesian estimation makes the results easier to understand due to its intuitive inference process (van de Schoot et al., 2013a). In the frequentist approach, a confidence interval is provided which shows that over an infinity of samples taken from the population, 95% of these contain the true population value. The interpretation of such an interval is somewhat counterintuitive, as it refers to samples rather than an actual parameter of interest. On the other hand, a Bayesian credibility interval indicates that there is a 95% chance for a parameter to lay within the limits of the interval. Taking this paper as an example, the credibility interval will reflect the most probable range of values for the correlation between the latent factors reflecting hedonic and eudaimonic pursuits of well-being. In other words, Bayesian approach focuses on the magnitude of the parameter for a provided dataset. Such information, in contrast to the traditional confidence intervals, is easier to understand and compare between groups.

In sum, the paper provides a practical application of Bayesian estimation, and aims at investigating some differences between the traditional frequentist approach to CFA with that of a Bayesian approach.

MATERIALS AND METHODS

PARTICIPANTS

In the Polish sample, 386 adults were surveyed, of whom 79% were female. Their age ranged from 18 to 29 years ($M = 21.26$, $SD = 1.75$). The data collection was conducted in two waves: first ($N = 197$) with the full 9-item version of the scale, and second ($N = 189$) with the short 8-item version (see the Supplementary Material for the list of items included in both versions). The assessment was based on a structured, anonymous questionnaire investigating a number of lifestyle-related variables (see Kaczmarek et al., 2013). Participation was on a voluntary basis and administration took place during the respondents' free time.

The English sample consisted of 429 North American Anglophone participants. The study involved undergraduates (75% of women) who completed the questionnaire as part of a 1-h screening survey (including measures submitted by a variety of researchers) used as a preliminary step before granting students access to various individual studies. Their age ranged from 18 to 30 years ($M = 19.19$, $SD = 1.92$).

INSTRUMENT

The HEMA scale is meant to assess motives for activities that can be divided into those that are eudaimonic (e.g., "seeking to develop the best in oneself") and those that are hedonic (e.g., "seeking pleasure"). It is underlined by Huta and Ryan (2010)—the authors of the scale—that this approach allows the distinguishing of hedonia and eudaimonia as forms of well-being pursuits from well-being products. It also offers the opportunity to study both motives as separate variables. Thus, the HEMA measures eudaimonia and hedonia in parallel terms, operationalizing both as orientations (Huta and Ryan, 2010).

The HEMA is a 9-item instrument comprising a hedonic motivation subscale (5 items) and a eudaimonic motivation subscale (4 items). Responses range from 1 (not at all) to 7 (very much). A back-translation procedure was used to translate the HEMA scale into Polish by two bilingual psychologists (the original English items and their Polish translation are presented in the Supplementary Table 1). During this process, one of the items ("seeking enjoyment"), originally belonging to the hedonic subscale, was identified to have different cultural connotations. In the English language, it reflects striving after pleasant experiences, while in the Polish version, it may have been perceived as reflecting goal achievement, rather than being specific to hedonia or eudaimonia. Such disparity could be expected due to the existence of different well-being definitions across nations (e.g., Wierzbicka, 2004). In this situation a partial MI that leaves out the problematic item could have been employed. However, this would undermine the interpretation of estimated factor correlations (the item reflected the hedonic construct in the English version only). Therefore, in this article we tested a possibility to use the 8-item version of the scale in both language groups.

ANALYSIS

The analysis was conducted in three stages: (1) the dimensional structure of the scale was established through CFA separately for the two national samples, (2) the MI was tested between the countries, and finally (3) the differences between the latent factors' correlations were tested. The research question focuses on the construct validity of the scale, therefore a metric invariance (equality of factor loadings) was of main interest (Byrne, 2012). This type of MI indicates whether respondents across groups attribute the same meaning to the latent construct under the study (van de Schoot et al., 2012). In other words, indicators that are central to the construct in one national group, are also central in the other (Selig et al., 2008). It is assumed here that the similar pattern of item intercepts (scalar invariance) is not required for a meaningful comparison of factor covariances, even though this claim can be considered controversial by some researchers (for discussion see Byrne and van de Vijver, 2010). Although of secondary interest, further analyses of the scalar invariance (i.e., invariance of observed variables' intercepts) were also conducted, and differences between intercepts were tested. This allowed for a better illustration of the functioning of the scale in the two national groups.

All the analyses were performed using Mplus 7.11 (Muthén and Muthén, 1998–2012). For the traditional analyses, ML parameter estimates with standard errors and a chi-square test statistic robust to non-normality was used (MLR, see Muthén and Muthén, 1998–2012). When ML estimation was employed, for the evaluation of a model the following fit indices were used with the respective cut-off values as proposed by Schweizer (2010);

χ^2, normed χ^2 (NC, with values below 3 indicating an acceptable fit and below 2 a good fit), CFI and TLI (acceptable model fit when higher than 0.90, good fit when higher than 0.95), RMSEA (acceptable fit when lower than 0.08, good fit when lower than 0.05) and SRMR (expected to stay below 0.10). Chi-square difference test (using the Satorra-Bentler scaled chi-square), AIC and BIC values were employed to compare models. In the Bayesian analyses, two indicators of a model fit were interpreted: (1) the posterior predictive *p*-value (PPP, good fit when equal to or higher than 0.05), and (2) the 95% confidence interval of the replicated chi-square value (which was expected to include zero; for details please refer to Muthén and Asparouhov, 2012). We additionally used the deviance information criterion to compare the model (DIC; Spiegelhalter et al., 2002). Model estimation was performed with maximum 500,000 and minimum 20,000 iterations using the Markov chain Monte Carlo (MCMC) algorithm (Muthén and Asparouhov, 2012). MCMC convergence criterion using potential scale reduction (PSR) was set to 0.01 (Gelman and Rubin, 1992). The alignment method for the approximate MI was not employed since it is not yet available for models with cross-loadings (Muthén and Asparouhov, 2014). The data and all Mplus output files are available in the Supplementary Materials.

RESULTS

FACTOR STRUCTURE

In the first step of the analysis a latent structure of the scale had to be established. Based on the theoretical assumptions underlying the HEMA scale, a model separating the hedonic and the eudaimonic factor was expected. Previous exploratory analyses revealed the existence of such two-factor structure (Huta and Ryan, 2010; Anić, 2014). However, a recent confirmatory analysis of the Japanese version of the HEMA scale revealed that a three-factor structure is a better representation of the scale's structure (Asano et al., 2014). Therefore, our goals were to (1) determine the factor structure of the scale, and (2) validate the performance of the short 8-item instrument. In order to do so, each national sample was divided into two groups. In the Polish sample the groups were formed according to the waves of the data collection (in the first wave the 9-item scale was administered, in the second wave the 8-item instrument was used). In the English sample participants were divided at random into two groups, and for the second group the "seeking enjoyment" item was removed from the analyses. Then, the one, two and three-factor solutions were tested in four samples. The analysis was begun with the traditional frequentist approach, followed by the Bayesian estimation.

The results acknowledged that the three-factor model was a better solution (see **Table 1**, syntaxes 1–7 included in the Supplementary Materials). In both national groups, and for both the short and full versions of the scale, splitting the hedonic factor into two components would notably improve the fit. Due to both the theoretical and empirical plausibility of such a distinction, we decided to continue the analyses with the three-factor structure. The proposed three-factor model categorized the hedonic items into a comfort ("seeking to take it easy"; "seeking relaxation") and a pleasure group ("seeking fun"; "seeking pleasure"; "seeking enjoyment" in the full version of the scale). The confirmatory

Table 1 | The confirmatory factor analyses using maximum likelihood estimation with robust standard errors (ML).

	NC	χ^2	*df*	*p*	RMSEA	CFI	TLI	SRMR	AIC	BIC
ENGLISH										
9-items (*N* = *223*)										
1 factor	12.09	326.50	27	< 0.001	0.22	0.55	0.40	0.15	6634	6726
2 factors	5.49	142.67	26	< 0.001	0.14	0.82	0.76	0.09	6430	6526
3 factors	2.13	55.44	26	< 0.001	0.08	0.95	0.93	0.05	6334	6436
8-items (*N* = *206*)										
1 factor	8.56	171.15	20	< 0.001	0.19	0.60	0.45	0.13	5682	5762
2 factors	4.86	92.27	19	< 0.001	0.14	0.81	0.72	0.09	5581	5664
3 factors	1.84	31.36	17	0.02	0.06	0.96	0.94	0.05	5507	5597
8-items, full sample (*N* = *429*)										
3 factors	3.36	57.16	17	< 0.001	0.07	0.95	0.92	0.05	11,303	11,412
POLISH										
9-items (*N* = *197*)										
1 factor	4.25	114.90	27	< 0.001	0.13	0.81	0.74	0.10	4967	5056
2 factors	2.39	62.17	26	< 0.001	0.08	0.92	0.89	0.07	4907	4999
3 factors	1.74	41.82	24	0.01	0.06	0.96	0.94	0.06	4889	4987
8-items (*N* = *189*)										
1 factor	6.47	129.46	20	< 0.001	0.17	0.65	0.51	0.12	4525	4603
2 factors	3.46	65.73	19	< 0.001	0.11	0.85	0.78	0.07	4451	4532
3 factors	2.39	40.76	17	0.001	0.09	0.92	0.87	0.06	4427	4515
8-items, full sample (*N* = *386*)										
3 factors	2.50	42.60	17	< 0.001	0.06	0.96	0.94	0.05	8878	8985

procedure verified this model by revealing its acceptable fit in both the English and Polish samples (**Table 1**). In this traditional approach to the CFA, no cross-loadings between items were allowed.

In terms of the short vs. the full version of the scale, the results were somewhat inconclusive. In the English sample the short version (with one item excluded from the analysis) fitted the data slightly better. In the Polish sample, however, the fit was worse when the 8-item version of the scale was administered. To check whether these differences represented the specific variability of a group, rather than a general tendency, we have retested the chosen three-factor model on the full English and Polish samples (using the short version of the scale). This resulted with acceptable fit in both national samples leading to a conclusion that the 8-item instrument produces similar factor structure to the one detected for the full version of the scale.

Bayesian CFA

The factor structure of the HEMA scale was then re-tested using Bayesian estimation (see **Table 2** for results, and syntaxes 8–11). First, the noninformative prior distribution was specified using the default prior settings available in Mplus (Muthén and Muthén, 1998–2012). Therefore, no previous knowledge was imposed, meaning that every value of a parameter was equally likely to occur (van de Schoot et al., 2013a). In this case, cross-loadings were fixed to zero, just as in the ML estimated models. With this specification, the two-factor and three-factor models were tested. The one-factor model was omitted for clarity, and due to its very poor fit as revealed in the previous analysis.

In almost all situations neither the two- nor three-factor model resulted in a satisfactory fit (i.e., the PPP was significant, and the 95% CI of the replicated chi-square values did not include zero). The only exception was the 8-item scale in the English group, where the three-factor model resulted in close to acceptable fit (i.e., the PPP was significant, but the 95% CI included zero). We then changed the requirements of the model so that cross-loadings would be approximately zero rather than exactly zero (syntaxes 12–16). Using small-variance priors (prior mean = 0; prior variance = 0.01) all cross-loadings were restricted to having a value ranging from −0.2 to 0.2 (for more choices please refer to Muthén and Asparouhov, 2012, p. 316). The goal of this strategy was to allow for cross-loadings, yet at the same time keep them small and statistically insignificant. This resulted in an improvement of the three-factor model fit, yet did not help in the case of the two-factor model (see **Table 2**). Thus, it was again concluded that the three-factor model fit the data better and the analyses were continued employing this structure.

Thanks to the use of weakly informative priors for cross-loadings the results of the Bayesian CFA provided some interesting insights into the performance of the short and full versions of the scale. In the English sample both the 9- and 8-item versions resulted with a good fit when cross-loadings were introduced in the three-factor solution (PPP was accordingly 0.34 and 0.17). In the Polish sample the short version of the scale responded with improvement into an almost acceptable model fit (PPP = 0.04; 95% CI included zero). However, when cross-loadings were allowed in the full version the model yielded a satisfactory fit also for the two-factor solution (PPP = 0.28). In both two- and three-factor models cross-loadings for the problematic "seeking enjoyment" item were large enough to become significant (see the Supplementary Table 3 for details). This suggested that in the Polish sample the 8-item version of the scale represents the measured constructs in a more clear way. Finally, the analyses conducted on the full English and Polish samples confirmed the fit of the three-factor model with small cross-loadings.

Table 2 | The confirmatory factor analyses using Bayesian estimation.

	#fp	2.5% pp	97.5% pp	PPP	DIC
ENGLISH					
9-items (*N* = *223*)					
2 factors NI	28	107.09	157.27	< 0.01	6432
2 factors CL	37	92.71	147.25	< 0.01	6424
3 factors NI	30	6.35	60.21	< 0.01	6335
3 factors CL	48	−20.42	34.94	0.34	6314
8-items (*N* = *206*)					
2 factors NI	25	74.39	123.23	< 0.01	5582
2 factors CL	33	49.83	107.29	< 0.01	5564
3 factors NI	27	−0.25	47.65	0.03	5509
3 factors CL	43	−14.84	39.37	0.17	5501
8-items, full sample (*N* = *429*)					
3 factors CL	43	−16.71	35.67	0.23	11,265
POLISH					
9-items (*N* = *197*)					
2 factors NI	28	19.54	70.05	< 0.01	4909
2 factors CL	37	−20.30	37.21	0.28	4877
3 factors NI	30	0.27	55.22	0.02	4895
3 factors CL	48	−24.75	33.42	0.38	4875
8-items (*N* = *189*)					
2 factors NI	25	28.94	78.39	< 0.01	4451
2 factors CL	33	19.75	72.00	< 0.01	4447
3 factors NI	27	6.25	53.19	< 0.01	4430
3 factors CL	43	−3.61	54.19	0.04	4427
8-items, full sample (*N* = *386*)					
3 factors CL	43	−12.27	37.86	0.14	8864

NI, Noninformative priors; CL, Informative priors on cross-loadings have a zero mean and a variance of 0.01.

MEASUREMENT INVARIANCE

In the second step of the analysis, a series of multi-group CFA were executed in order to test the MI between the Polish and the English versions of the HEMA scale (syntaxes 17–21). Stepwise procedures were employed, where the analysis begins with the least restricted solution and subsequent models with increasingly restrictive constraints are evaluated (Brown, 2006). Comparisons were performed with a corrected chi-square differences test due to the fact that the analyses were based on a robust maximum likelihood method (MLR; Muthén and Muthén, 1998–2012). In order to identify the model the factor variance was fixed to 1 in one group only, and in the other group the equality constraints were placed on the factor loadings while a factor variance was estimated (Yoon and Millsap, 2007). This method minimizes problems

Table 3 | The measurement invariance analyses using ML.

	NC	χ^2	*df*	RMSEA	CFI	TLI	SRMR	$\Delta\chi^{2\ adj}$	Δdf	Δp
Configural	2.95	100.19	34	0.07	0.96	0.93	0.05	–	–	–
Metric	2.92	113.73	39	0.07	0.95	0.93	0.06	13.61	5	0.02
Partial[a]	2.78	105.64	38	0.07	0.95	0.93	0.06	5.98	4	0.20
Scalar	5.55	238.86	43	0.11	0.87	0.83	0.08	179.32	5	< 0.001
Partial[b]	2.75	107.327	39	0.07	0.95	0.93	0.06	1.38	1	0.24

[a] *Free factor loading of item 1 (relaxation), the partial metric model is compared to the configural model.*

[b] *Free intercepts of items 4 (pleasure), 3 (do what you believe), 2 (learn, develop skills), and 1 (relaxation), the partial scalar model is compared to the partial metric model.*

caused by commonly used solutions such as constraining the first factor loading to one (Bauer and Hussong, 2009).

With a configural invariance model (syntax 17) established across the two language versions of the HEMA, the next step was to test for metric invariance (see **Table 3**). A model constraining factor loadings to being equivalent across the language versions (syntax 18) fitted slightly worse than a configural model. Then a partial metric invariance model (syntax 19) was tested, where one factor loading was allowed to vary across groups ("seeking relaxation"). This specification represented the data well. Accordingly, the construct validity of the scale was confirmed across the national samples enabling a meaningful comparison between the factor covariances. Finally, a scalar invariance was tested (syntax 20). The results were not supportive for the scalar invariance indicating that the intercepts were not equal across the samples. Partial scalar model (syntax 21) could have been established only when half of the intercepts were allowed to vary. We have therefore retreated to the partial metric MI model and this one was applied for the final step of the analyses (see **Table 5** for standardized results of the partial metric MI model).

Approximate MI

Then, the MI across groups was re-tested with the Bayesian estimator (see **Table 4**). We have continued with the model where weakly informative priors on the cross-loadings were used (i.e. small cross-loadings were allowed, see syntaxes 22–26). The configural model (syntax 22) fitted the data well as expected, given the previously confirmed stability of the three-factor model across the national groups. The full metric model (syntax 23) resulted in an acceptable fit, yet the PPP-value was still quite low (0.057). We have proceeded to establish and approximate metric invariance model (syntax 24), resulting in a slightly better fit (higher PPP-value, 0.078), but the difference in DIC was small (equals 2.1). Therefore, we have decided to employ the metric invariance model, not the approximate one, in the further analysis. The standardized factor loadings and intercepts estimated for this model are presented in **Table 5**. Lastly, the scalar invariance (syntax 25) was tested, but the model did not represent the data well. Therefore, the strict equality assumption between the intercepts was released, and the approximate scalar MI (syntax 26) was implemented (Muthén and Asparouhov, 2013). Allowing all the intercepts to be at least approximately equal (prior mean = 0; prior variance = 0.01) did not help to improve the fit. In fact neither of the methods used (the partial invariance with ML or the approximate invariance with Bayes) supported scalar invariance. This suggests that the problem of non-invariant intercepts is not limited to a particular item(s), and that all the items vary to an extent that cannot be disregarded (only for the item 5 credibility intervals for the intercepts overlapped) The intercepts in the Polish sample were higher than the ones of the English sample (see **Table 5**). Thus, the comparison of factor means would not be possible. Yet, in order to compare factor covariances the metric model could be used (Byrne and van de Vijver, 2010).

Table 4 | The measurement invariance analyses using Bayesian estimation.

BSEM	#fp	2.5% pp	97.5% pp	PPP	DIC
Configural	86	−13.57	61.47	0.102	20,133
Metric	81	−10.53	67.75	0.057	20,131
Metric approximate[a]	89	−10.97	60.773	0.078	20,129
Scalar	76	21.63	91.06	0.001	20,154
Scalar approximate[a]	84	11.76	86.44	0.005	20,149

[a] *Informative priors on differences between groups have a zero mean and a variance of 0.01.*

CORRELATION BETWEEN HEDONIA AND EUDAIMONIA

The third and final step of the analysis was to quantify and compare the covariances between the latent factors representing hedonic and eudaimonic pursuits of well-being (syntaxes 27–28). To do so, the MODEL CONSTRAINT command was included in the Mplus code (Muthén and Muthén, 1998–2012). This function allowed us to create a set of new parameters representing the differences between the estimated factor covariances across and within the national groups. Mplus provided us with confidence intervals (using the Delta method standard errors and *z*-test for ML estimation) or credibility intervals (for Bayesian estimation) for the newly defined parameters. As a result, the differences between the factor covariances were tested for statistical significance. We continued to use partial metric invariance model for MI estimation (syntax 27), and metric invariance model with small cross-loadings specified with weakly informative priors for Bayes (syntax 28).

Table 6 presents the factor correlations (standardized covariances) for the two national groups and the two estimation methods, marked for the significant differences between and

Table 5 | Standardized factor loadings and intercepts for the metric invariance model.

	ML				Bayes			
	Polish		English		Polish		English	
	FL	Int	FL	Int	FL	Int	FL	Int
HEDONIC PLEASURE								
Item 4 (pleasure)	0.82	5.31	0.80	3.52	0.76	5.28	0.73	3.51
Item 8 (fun)	0.68	3.94	0.80	3.64	0.74	3.88	0.86	3.65
HEDONIC COMFORT								
Item 1 (relaxation)	0.93	4.46	0.70	2.70	0.88	4.40	0.82	2.72
Item 6 (easy)	0.75	4.32	0.90	2.65	0.81	4.32	0.77	2.63
EUDAIMONIC FACTOR								
Item 2 (learn, develop skills)	0.50	5.37	0.62	3.77	0.55	5.35	0.67	3.74
Item 3 (do what you believe)	0.51	4.81	0.61	3.05	0.51	4.81	0.60	3.02
Item 5 (pursue excellence)	0.57	4.58	0.80	3.80	0.59	4.52	0.82	3.80
Item 7 (use the best in yourself)	0.60	4.83	0.80	3.74	0.56	4.78	0.74	3.74

FL, Factor loadings; Int, Intercepts. Partial metric invariance model in the ML estimation. Cross-loadings omitted for clarity.

Table 6 | Correlations between latent factors of hedonia and eudaimonia as estimated with ML and Bayes.

	ML (95% *CI*)		Bayes (BCI)	
	Polish	English	Polish	English
Hedonic pleasure with hedonic comfort	0.82^{1} (0.74; 0.91)	0.46^{1} (0.32; 0.59)	0.82^{1} (0.72; 0.90)	0.47^{1} (0.28; 0.62)
Hedonic pleasure with eudaimonic	0.29^{2} (0.12; 0.46)	0.54^{1} (0.42; 0.67)	0.26^{2a} (0.01; 0.48)	0.50^{1a} (0.31; 0.66)
Hedonic comfort with eudaimonic	0.18^{2a} (0.03; 0.33)	0.09^{1a} (− 0.01; 0.25)	0.17^{2} (− 0.06; 0.38)	0.16^{2} (− 0.03; 0.34)

Correlations marked with superscript letter "a" differ between national groups. Correlations within one column not sharing the same superscript number differ within national groups. CI, Confidence interval; BCI, Bayesian credibility interval.

within groups. Both the traditional and Bayesian procedures revealed similar latent factor correlation patterns within each of the national groups. In the Polish sample the connection between the two hedonic factors was found to be the strongest (significantly stronger than each of the other two correlations), while the correlations between the hedonic factors and the eudaimonic factor were rather weak. In the English sample the links connecting the hedonic factors, and the hedonic pleasure with the eudaimonic factor were moderate. The correlation between the hedonic comfort and the eudaimonic factor was weak and mostly insignificant (only in the Polish group with ML $p = 0.02$). It was weaker than each of the other two correlations in all cases except from ML estimation in the English group. Between group differences were found in the connection between hedonic comfort with the eudaimonic factor (with ML) or in the hedonic pleasure with eudaimonic factor (with Bayes). Even though the correlation between the hedonic factors was stronger in the Polish group than the English group, neither ML nor Bayes found this difference significant.

DISCUSSION

The aim of this paper was to describe the structure of the HEMA scale and its performance across two different nations. Stepwise analyses were conducted to establish a factor structure of the scale, revealing three correlated factors: two hedonic and one eudaimonic. The eudaimonic factor reflected the pursuit for excellence. The hedonic factors include items reflecting the pursuit of affective states and were divided into a comfort factor and a pleasure factor.

Among the two hedonic factors, the one reflecting pleasure was closer to the eudaimonic factor than was the hedonic comfort factor. This pattern was relatively stable across the national groups. Thus, seeking excellence seems to feel more like pleasure and fun, than like being relaxed and at ease. In fact, some researchers include the enjoyment from activities representing the pursuit of excellence into their definition of eudaimonia (Waterman et al., 2010). At the same time, both the hedonic factors were strongly correlated, indicating that these items roughly occupy the same area of the affective landscape. Splitting hedonia into the two components shed more light on unclear previous results regarding its connection to eudaimonia (Huta and Waterman, 2013). This division may be valid for further research, yet building a dedicated scale to assess hedonic comfort and pleasure is recommended.

This paper is a first attempt at a systematic examination of multigroup stability of the HEMA scale components. Future work should address several issues not answered in this study. First of all, the conclusions of this paper are based on the models that fit

acceptably well, yet are not perfect. More national samples should be taken into consideration to further justify the cross-national stability of a well-being assessment, including both hedonic and eudaimonic constructs. Secondly, representative samples are preferred in order to avoid possible sampling errors. Especially the lack of gender balance is an important limitation of this study. Thirdly, the lack of scalar invariance revealed in this study should be examined more closely. The differences between intercepts are not surprising when comparing across nations, as they might occur as a result of different norms of socially-acceptable levels for expressions of hedonia and eudaimonia (e.g., Diener, 2000). Detailed analyses of this phenomenon were not within the scope of this paper, but remain an interesting issue. Finally, the division of hedonic comfort and pleasure needs further attention, and a more detailed assessment of those hedonic components could be considered.

In summary, this paper revealed a similar pattern of correlations between the trait-level pursuits of hedonic and eudaimonic elements across the two national samples. Further cross-national studies are needed to confirm the existence of this pattern, as well as to explain the differences in the items' intercepts across the groups (the lack of scalar invariance). It is hoped that using the HEMA scale in various language versions and across different national groups has the potential to substantially advance the knowledge on hedonic and eudaimonic components of well-being.

THE PERFORMANCE OF THE BAYESIAN ESTIMATION

Bayesian estimation was employed in this study due to its fundamental advantages over the traditional frequentist approach (e.g., Muthén and Asparouhov, 2012; van de Schoot et al., 2013a; Zyphur and Oswald, 2013). It was expected that specifying weakly informative priors would help us to better assess the differences between groups, and the intuitive inference process would provide a simpler interpretation of the factor correlations. Several points regarding the fulfillment of those expectations are discussed here.

Firstly, Bayesian estimation reported a misfit of the model when strong assumptions of exact zero were imposed. This is interesting given that the estimation based on the ML method reported an acceptable fit. The reason for this lies within the definition of a model fit used in Bayesian estimation. The posterior predictive checking assesses how well a model is specified from the viewpoint of predictive accuracy (how well it predicts the data). Thus, any discrepancies are detected between the values generated by a model and the observed data, suggesting that the model could be improved (van de Schoot et al., 2013a).

Consequently, in Bayesian estimation replacing the exact zero assumption with an approximate zero improved the fit significantly, leading to the acceptance of the model with small cross-loadings. In fact Bayes arrived at a similar outcome to that of the traditional estimation, yet using a longer route. This detour, however, was much more informative. While in the ML approach the CFA is not able to provide information about the reason of model misfit, Bayesian modeling gives more hints about it. Including small priors for cross-loadings (or residual covariances which is also possible, see Muthén and Asparouhov, 2012) helps in verifying why the model does not represent the data well. Yet, it should be underlined that when large discrepancies were observed, such as when scalar invariance was imposed, changing the exact zero assumption into the approximate one (in this case by specifying an approximate MI) did not help in achieving a satisfactory fit. Clearly, according to both ML and Bayes, the differences between the items' intercepts were too big for scalar invariance to be established. This shows that the Bayesian approach can be more informative than ML only when the models are already fairly well specified. Indeed, this method is advised for analyses with a small number of groups, continuous variables and close-to-invariant models (van de Schoot et al., 2013b).

Finally, taking into account the small cross-loadings might have been the reason why the Bayesian estimation did not discover any differences between factor loadings (allowing for full metric invariance). Interestingly, for both methods the estimated factor correlations were almost identical, even though ML used only the partial metric MI model. In this case including small cross-loadings did not influence the structural parameters of the model. In fact, it helped in establishing the metric invariance. This might suggest that non-invariant cross-loadings (not included in the traditionally estimated metric MI model) could actually be the reason for its misfit. Such possibility opens up an interesting discussion, but simulation studies are needed to better understand the role of small and sample specific cross-loadings in multigroup MI analyses.

To summarize, Bayesian estimation can be a recommended approach to MI analyses when (1) small differences between groups are expected and the size of those differences should be estimated, and (2) when structural parameters are of interest (e.g., factor covariances) and a researcher would like to be provided with easy to interpret credibility intervals for such parameters.

ACKNOWLEDGMENT

This research was supported in part by a grant from the Research Council of Norway (Yggdrasil programme 2012, project number 210845).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00984/abstract

REFERENCES

Anić, P. (2014). Hedonic and eudaimonic motives for favourite leisure activities. *Primenjena Psihologija* 7, 5–21.

Asano, R., Igarashi, T., and Tsukamoto, S. (2014). Hedonic and eudaimonic motives for activities (HEMA) in Japan: the pursuit of well-being. *Jpn. J. Psychol.* 85, 69–79. doi: 10.4992/jjpsy.85.69

Bauer, D. J., and Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychol. Methods* 14, 101–125. doi: 10.1037/a0015583

Bauer, J., Park, S., Montoya, R. M., and Wayment, H. (2014). Growth motivation toward two paths of eudaimonic self-development. *J. Happiness Stud.* doi: 10.1007/s10902-014-9504-9. [Epub ahead of print].

Berridge, K., and Kringelbach, M. (2011). Building a neuroscience of pleasure and well-being. *Psychol. Well Being* 1:3. doi: 10.1186/2211-1522-1-3

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *J. Organ. Behav.* 16, 201–213. doi: 10.1002/job.4030160303

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: The Guilford Press.

Byrne, B. (2012). *Structural Equation Modeling with Mplus: Basic Concepts, Applications, and Programming*. New York, NY: Taylor & Francis Group.

Byrne, B., Shavelson, R., and Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Byrne, B., and van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: addressing the issue of nonequivalence. *Int. J. Test.* 10, 107–132. doi: 10.1080/15305051003637306

Cabanac, M. (2010). *The Fifth Influence: Or, The Dialectics of Pleasure*. Bloomington, IN: iUniverse.

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., and Schwartz, S. H. (2014). Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: a cross-country illustration with a new scale to measure 19 human values. *Front. Psychol.* 5:982. doi: 10.3389/fpsyg.2014.00982

Deci, E. L., and Ryan, R. M. (2008). Hedonia, eudaimonia, and well-being: an introduction. *J. Happiness Stud.* 9, 1–11. doi: 10.1007/s10902-006-9018-1

Diener, E. (1984). Subjective well-being. *Psychol. Bull.* 95, 542–575. doi: 10.1037/0033-2909.95.3.542

Diener, E. (2000). Subjective well-being: the science of happiness and a proposal for a national index. *Am. Psychol.* 55, 34–43. doi: 10.1037/0003-066X.55.1.34

Diener, E., Scollon, C. N., and Lucas, R. E. (2009). "The evolving concept of subjective well-being: the multifaceted nature of happiness," in *Assessing Well-being: The Collected Works of Ed Diener*, ed E. Diener (Dordrecht, NL: Springer), 67–100.

Elliot, A. J. (2008). *Handbook of Approach and Avoidance Motivation*. New York, NY: Psychology Press.

Feldman, F. (2004). *Pleasure and the Good Life: Concerning the Nature, Varieties, and Plausibility of Hedonism*. Oxford, UK: Clarendon Press.

Gelman, A., and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Golay, P., Reverte, I., Rossier, J., Favez, N., and Lecerf, T. (2013). Further insights on the French WISC-IV factor structure through Bayesian structural equation modeling. *Psychol. Assess.* 25, 496–508. doi: 10.1037/a0030676

Henderson, L. W., Knight, T., and Richardson, B. (2013). An exploration of the well-being benefits of hedonic and eudaimonic behaviour. *J. Posit. Psychol.* 8, 322–336. doi: 10.1080/17439760.2013.803596

Huta, V. (2013). "Eudaimonia," in *The Oxford Handbook of Happiness*, eds S. David, I. Boniwell, and A. C. Ayers (Oxford, UK: Oxford University Press), 201–213.

Huta, V., and Ryan, R. M. (2010). Pursuing pleasure or virtue: the differential and overlapping well-being benefits of hedonic and eudaimonic motives. *J. Happiness Stud.* 11, 735–762. doi: 10.1007/s10902-009-9171-4

Huta, V., and Waterman, A. (2013). Eudaimonia and its distinction from hedonia: developing a classification and terminology for understanding conceptual and operational definitions. *J. Happiness Stud.* doi: 10.1007/s10902-013-9485-0. [Epub ahead of print].

Kaczmarek, L. D., Kashdan, T. B., Kleiman, E. M., Baczkowski, B., Enko, J., Siebers, A., et al. (2013). Who self-initiates gratitude interventions in daily life? An examination of intentions, curiosity, depressive symptoms, and life satisfaction. *Pers. Individ. Diff.* 55, 805–810. doi: 10.1016/j.paid.2013.06.013

Kahneman, D. (1999). "Objective happiness," in *Well-being: The Foundations of Hedonic Psychology*, eds D. Kahneman, E. Diener, and N. Schwarz (New York, NY: Russell Sage Foundation), 3–25.

Kahneman, D. (2000). "Experienced utility and objective happiness: a moment-based approach," in *Choices, Values and Frames*, eds D. Kahneman and A. Tversky (New York, NY: Cambridge University Press and the Russell Sage Foundation), 673–692.

Keyes, C., Shmotkin, D., and Ryff, C. (2002). Optimizing well-being: the empirical encounter of two traditions. *J. Pers. Soc. Psychol.* 82, 1007–1022. doi: 10.1037//0022-3514.82.6.1007

Keyes, C. L. M., and Annas, J. (2009). Feeling good and functioning well: distinctive concepts in ancient philosophy and contemporary science. *J. Posit. Psychol.* 4, 197–201. doi: 10.1080/17439760902844228

Kopperud, K. H., and Vittersø, J. (2008). Distinctions between hedonic and eudaimonic well-being: results from a day reconstruction study among Norwegian jobholders. *J. Posit. Psychol.* 3, 174–181. doi: 10.1080/17439760801999420

Meredith, W., and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Med. Care* 44(11 Suppl. 3), S69–S77. doi: 10.1097/01.mlr.0000245438.73837.89

Millsap, R. E., and Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychol. Methods* 9, 93–115. doi: 10.1037/1082-989X.9.1.93

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

Muthén, B., and Asparouhov, T. (2013). *BSEM Measurement Invariance Analysis. Mplus Web Notes: No. 17.* Available online at: www.statmodel.com

Muthén, B., and Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Front. Psychol.* 5:978. doi: 10.3389/fpsyg.2014.00978

Muthén, L. K., and Muthén, B. (1998–2012). *Mplus User's Guide*. 7th Edn. Los Angeles, CA: Muthén and Muthén.

Oishi, S., Graham, J., Kesebir, S., and Galinha, I. C. (2013). Concepts of happiness across time and cultures. *Pers. Soc. Psychol. Bull.* 39, 559–577. doi: 10.1177/0146167213480042

Proctor, C., Tweed, R., and Morris, D. (2014). The naturally emerging structure of well-being among young adults: "Big Two" or other framework? *J. Happiness Stud.* doi: 10.1007/s10902-014-9507-6. [Epub ahead of print].

Raykov, T., Marcoulides, G. A., and Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations: a critical view and refocus. *Educ. Psychol. Meas.* 72, 954–974. doi: 10.1177/0013164412441607

Ryan, R. M., and Deci, E. D. (2001). On happiness and human potentials: a review of research on hedonic and eudaimonic well-being. *Annu. Rev. Psychol.* 52, 141–166. doi: 10.1146/annurev.psych.52.1.141

Ryan, R. M., Huta, V., and Deci, E. L. (2013). "Living well: a self-determination theory perspective on eudaimonia," in *The Exploration of Happiness: Present and Future Perspectives*, ed A. Delle Fave (New York, NY: Springer Science), 117–139.

Ryff, C. D. (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *J. Pers. Soc. Psychol.* 57, 1069–1081. doi: 10.1037/0022-3514.57.6.1069

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

Schweizer, K. (2010). Some guidelines concerning the modeling of traits and abilities in test construction. *Eur. J. Psychol. Assess.* 26, 1–2. doi: 10.1027/1015-5759/a000001

Scitovsky, T. (1976). *The Joyless Economy: The Psychology of Human Satisfaction*. New York, NY: Oxford University Press.

Selig, J. P., Card, N. A., and Little, T. D. (2008). "Latent variable structural equation modeling in cross-cultural research: multigroup and multilevel approaches," in *Multilevel Analysis of Individuals and Cultures*, eds. J. R. van de Vijver, D. A. van Hemert, and Y. H. Poortinga (New York, NY: Taylor & Francis), 93–119.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* 64, 583–639. doi: 10.1111/1467-9868.00353

Tännsjö, T. (2007). Narrow hedonism. *J. Happiness Stud.* 8, 79–98. doi: 10.1007/s10902-006-9005-6

Tatarkiewicz, W. (1976). *Analysis of Happiness.* The Hague: M. Nijhoff and Polish Scientific Publishers.

Uchida, Y., Takahashi, Y., and Kawahara, K. (2014). Changes in hedonic and eudaimonic well-being after a severe nationwide disaster: the case of the great east Japan earthquake. *J. Happiness Stud.* 15, 207–221. doi: 10.1007/s10902-013-9463-6

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., and van Aken, M. A. G. (2013a). A gentle introduction to Bayesian analysis: applications to developmental research. *Child Dev.* 85, 842–860. doi: 10.1111/cdev.12169

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthén, B. (2013b). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *Eur. J. Dev. Psychol.* 9, 486–492. doi: 10.1080/17405629.2012.686740

Vittersø, J. (2013). "Functional well-being: happiness as feelings, evaluations and functioning," in *The Oxford Handbook of Happiness*, eds S. David, I. Boniwell, and A. C. Ayers (Oxford, UK: Oxford University Press), 227–244

Waterman, A. S. (1993). Two conceptions of happiness: contrasts of personal expressiveness (eudaimonia) and hedonic enjoyment. *J. Pers. Soc. Psychol.* 64, 678–691. doi: 10.1037/0022-3514.64.4.678

Waterman, A. S., Schwartz, S. J., Zamboanga, B. L., Ravert, R. D., Williams, M. K., Agocha, V. B., et al. (2010). The Questionnaire for eudaimonic Well-Being: psychometric properties, demographic comparisons, and evidence of validity. *J. Posit. Psychol.* 5, 41–61. doi: 10.1080/17439760903435208

Wierzbicka, A. (2004). "Happiness" in cross-linguistic and cross-cultural perspective. *Daedalus* 34, 34–43. doi: 10.1162/001152604323049370

Yoon, M., and Millsap, R. E. (2007). Detecting violations of factorial invariance using data- based specification searches: a Monte Carlo study. *Struct. Equ. Model.* 14, 435–463. doi: 10.1080/10705510701301677

Zyphur, M., and Oswald, F. (2013). Bayesian estimation and inference: a user's guide. *J. Manage.* doi: 10.1177/0149206313501200. [Epub ahead of print].

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 April 2014; accepted: 19 August 2014; published online: 08 September 2014.

Citation: Bujacz A, Vittersø J, Huta V and Kaczmarek LD (2014) Measuring hedonia and eudaimonia as motives for activities: cross-national investigation through traditional and Bayesian structural equation modeling. Front. Psychol. ***5****:984. doi: 10.3389/fpsyg.2014.00984*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 19 September 2014
doi: 10.3389/fpsyg.2014.00983

An approximate measurement invariance approach to within-couple relationship quality

Carlo Chiorri[1,2]*, Thomas Day[3] and Lars-Erik Malmberg[3]

[1] *Department of Educational Sciences, University of Genoa, Genoa, Italy*
[2] *Psyche-Dendron Association, Italy*
[3] *Department of Education, University of Oxford, Oxford, UK*

Edited by:
Rens Van De Schoot, Utrecht University, Netherlands

Reviewed by:
Anne C. Black, Yale University, USA
Eldad Davidov, University of Zurich, Switzerland
Francesca Righetti, Vrije Universiteit Amsterdam, Netherlands

****Correspondence:***
Carlo Chiorri, Department of Educational Sciences, University of Genoa, Corso A. Podestà, 2-16128 Genoa, Italy
e-mail: carlo.chiorri@unige.it

This study aimed at demonstrating the usefulness and flexibility of the Bayesian structural equation modeling approximate measurement invariance (BSEM-AMI) approach to within-couple data. The substantive aim of the study was investigating partner differences in the perception of relationship quality (RQ) in a sample of intact couples ($n = 435$) drawn from the first sweep of the Millenium Cohort Study. Configural, weak and strong invariance models were tested using both maximum likelihood (ML) and BSEM approaches. As evidence of a lack of strong invariance was found, full and partial AMI models were specified, allowing nine different prior variances or "wiggle rooms." Although we could find an adequately fitting BSEM-AMI model allowing for approximate invariance of all the intercepts, the two-step approach proposed by Muthén and Asparouhov (2013b) for identifying problematic parameters and applying AMI only to them provided less biased results. Findings similar to the ML partial invariance model, led us to conclude that women reported a higher RQ than men. The results of this study highlight the need to inspect parameterization indeterminacy (or alignment) and support the efficacy of the two-step approach to BSEM-AMI.

Keywords: measurement invariance, Bayesian structural equation modeling, dyadic data, relationship quality, marital satisfaction

INTRODUCTION

In this study we present a worked example of the usefulness and flexibility of the recently developed Bayesian structural equation modeling approximate measurement invariance analysis (BSEM-AMI, Muthén and Asparouhov, 2013b) in addressing a common issue in relationship research, i.e., testing mean differences in partners' perception of relationship quality (RQ). This is a special case of gender differences testing, since the data from each individual are not unrelated to the data from every other individual in the study, as partners are nested within couples. This violates the assumption of independent errors and implies that, as we discuss in the Analytic strategy section, the unit of analysis has to be the couple, with women and men being different (but identifiable) raters of the same relationship. While our aim is not to draw definite conclusions about the long debated issue of partner differences in the perception of RQ, we offer to relationship researchers an example of a principled analytical approach to address it. For our didactic purposes we used partners' scores on a 7-item version of the Golombok-Rust Inventory of Marital State (Rust et al., 1986, 1990), which is included in the first sweep of University of London, Institute of Education, Centre for Longitudinal Studies (2012). As the psychometric properties of this short version have not been comprehensively tested, this study also provides evidence of its reliability, unidimensionality and partial measurement invariance across partners in intact couples.

THE SUBSTANTIVE FOCUS

Relationship quality (RQ), also referred to as marital quality, marital satisfaction or dissatisfaction, marital characterization, marital discord, marital conflict or relationship satisfaction is a key measure in family and developmental research. It has been linked to personal outcomes such as psychological and physical health of the partners, and with some crucial family outcomes such as domestic violence, poor parenting, and poor adjustment of children (Grych and Fincham, 2001; Fincham, 2003). Partners who lead a happy relationship are healthier (but see Robles et al., 2014), tend to communicate well with each other, are good-enough parents who raise their children authoritatively, and run less risk of marital breakdown (for more information see Section Introduction of the Supplementary Materials). Bradbury et al. (2000) highlighted the crucial role that RQ plays in sustaining individual and family-level well-being. A society benefits from such strong marital bonds that are formed and maintained as they provide a robust basis for bringing up children. Healthy children, parents and communities provide the rationale behind the need to "develop empirically defensible interventions for couples that prevent marital distress and divorce" (p. 964).

It is therefore important to understand with some precision how we can adequately measure the quality of relationships within couples. The use of reliable measures could assist practitioners (e.g., family therapists) to draw fine-tuned differences in partners' perceptions of their couple-life. In family research there are models in which the dependencies between partners

are modeled, such as the Actor-Partner-Interdependence-Model (APIM; Kenny, 1996; Kenny and Cook, 1999), as well as models in which within-couple agreement (correlations) or discrepancies (mean-level differences) are specified (e.g., Luo et al., 2008).

RQ has been operationally defined as a global evaluation of the relationship along several dimensions, including self-reported satisfaction with the relationship, attitudes toward one's partner, and levels of hostile and negative behavior (Robles et al., 2014). Numerous measures have been developed to assess it, from single-item measures (e.g., "How happy is your relationship with your partner, all things considered?" rated on a 7-point Likert-type scale, included in the National Child Development Study in UK), to multi-item measures, such as the Locke-Wallace Marital Adjustment Test (MAT, Locke and Wallace, 1959) or the Dyadic Adjustment Scale (DAS, Spanier, 1976, revised by Busby and Christensen, 1995, for use with distress and non-distressed couples) (for reviews see Child Trends, 2003; Bronte-Tinkew et al., 2004; Reynolds et al., 2014). In this study we focused on the Golombok-Rust Inventory of Marital State (GRIMS, Rust et al., 1986, 1990), whose 7-item shortened version was included in the first sweep (2003) of the Millenium Cohort Study (MCS). Rust et al. (1986) developed the GRIMS for use in couple counseling centers as a measure of change before and after treatment and was initially intended as a companion to the Golombok-Rust Inventory of Sexual Satisfaction (GRISS, Rust and Golombok, 1985). The original 28-item GRIMS allows to measure relationship change over time and to highlight relationship difficulties and focuses on two domains of the relationship, (1) shared interests, communication, sex, warmth, roles and decision making, and coping, and (2) beliefs about and attitudes toward relationships, behavior in the relationship and agreement with the partner. The 7-item GRIMS was developed to meet the need for a shorter measure of RQ to be included in the MCS questionnaire while retaining the content validity of the original version. Using archival data, the items to be included in the final version were chosen in order to (a) retain the framework of the original blueprint as much as possible; (b) obtain a similar number of positive and negative items and (c) achieve adequate corrected item-total correlations. The shortened scale was tested on a standardization sample of 266 individuals, and results showed good internal consistency (Cronbach's $\alpha = 0.86$ in both women and men), and no significant skewness or kurtosis (Rust, personal communication, 2014[1]). As for its criterion validity, previous findings from the MCS in relation to outcomes of RQ revealed that, women more satisfied with their relationship use less harsh discipline, parents in happier relationships spent more time with their children, women happier in relationships had children with higher British Ability Scale naming vocabulary scores and low RQ is linked to more behavioral problems (Jones, 2010).

One issue that has been extensively investigated in relation to RQ is whether there are gender differences, i.e., whether partners systematically experience different levels of RQ. As reported by Jackson et al. (2014) in their recent meta-analytic study, since Bernard (1972)'s seminal work it has long been assumed that women experience significantly less relationship satisfaction than men, but despite a number of studies supported this assumption, evidence for a lack of difference has also been provided (see Jackson et al., 2014 for a review). Jackson et al. (2014) concluded that there was a high average correlation of RQ scores between husband and wife pairs (0.51), and that wives were 51% less likely to be satisfied than their husbands only in couples undergoing marital therapy, whereas the difference was not significantly different from zero in community-based couples, especially in intact couples. This meta-analysis did not include studies that used the GRIMS. Rust et al. (1986, 1990) reported mixed results about gender differences in the 28-item scale scores: in the 1986 study, men obtained higher scores than women in the pilot sample and in clinical samples, while in the 1990 study men's scores were lower in clinical samples and equal in a sample of attendees at a general practitioners clinic. In the development study of the 7-item GRIMS, Rust (personal communication, 2014[1]) did not find significant gender differences in raw scores.

THE METHODOLOGICAL FOCUS

Both Jackson et al. (2014) and Rust et al. (1986, 1990) studies tested gender differences in RQ observed scores under the untested assumption that all the RQ measures included in the analysis showed measurement invariance across partners, i.e., the underlying measurement model of RQ measures was equivalent for both women and men. In particular, a crucial assumption in the comparison of RQ scores across spouses is, beyond the invariance of factor loadings, the invariance of item intercepts, i.e., whether the mean differences based on the latent construct are reflected in each of the individual items used to infer it. In other words, if the level of partner differences in RQ varies substantially from item to item for different items used to infer the construct, then the partner differences based on the corresponding latent construct should be considered idiosyncratic to the particular items used to infer RQ (i.e., differential item functioning). If this turns out to be true, results would suggest that conclusions about differences in RQ do not generalize over the set of items used in the instrument and the interpretation of latent mean comparisons among partners would be compromised (van de Schoot et al., 2012). In other words, even if partners rate the same items about the same relationship, their scores cannot be compared because the instrument does not measure the same construct in the same way.

Despite the wide use of RQ measures in surveys and research, very few studies have compared their factor structures across partners. One exception is South et al. (2009)'s study that demonstrated support for factorial invariance of the DAS across spouses. As pointed out by the authors, having established invariance of the DAS across gender, it can be concluded that any differences between men and women (as they found for dyadic consensus and affectional expression, with men scoring lower than women) can be interpreted as arising from actual differences in relationship adjustment, not that the instrument is measuring different constructs in the two groups. Besides, being able to reliably establish that there are systematic differences in scores between women and men would imply that different norms might be needed to interpret scores from either spouse. To the best of our knowledge no study has addressed this issue about the 7-item GRIMS, nor

[1]Rust, J. (2014). *Personal Communication*, July 14, 2014.

whether there are gender differences in scores. To this end, the aim of this study was to test its measurement invariance and investigate whether there are gender differences in its scores in a sample of intact couples from the first sweep of the MCS.

One frequent issue about measurement invariance is what to do when, after finding support for the ability of the a priori model to fit the data in each group without invariance constraints (*configural invariance*) and for the invariance of the factor loadings (*weak* or *metric invariance*), the model that imposes equality on item intercepts (*strong* or *scalar invariance*), does not fit, thus preventing a meaningful test of latent score differences. Muthén and Christoffersson (1981) suggested that it is possible to test invariance when only some of the parameters are invariant, and they termed this "partial" measurement invariance. Byrne et al. (1989) argued that full invariance is not necessary for performing further invariance tests and substantive analyses and proposed that mean comparisons would be meaningful if weak and strong invariance have been satisfied for at least two items per latent trait. Actually, the estimates of trait mean differences will be more accurately estimated with imposed partial invariance constraints, since the trait mean estimates are adjusted for the fact that only partial, not full, invariance characterizes the data: in other words, allowing the intercepts to vary automatically excludes the non-invariant items from the estimation of latent means (Cheung and Rensvold, 2000). Another approach to the problem, named Approximate Measurement Invariance (AMI), has been recently described by Muthén and Asparouhov (2012a, 2013b) and successfully implemented by van de Schoot et al. (2013). This method uses Bayesian structural equation models (BSEM) in which *exact* zero constraints can be replaced with *approximate* zero constraints based on substantive theories. In other words, differences in item intercepts that in confirmatory factor analysis would be constrained to be zero, under AMI can be estimated with some so-called "wiggle room" (Muthén and Asparouhov, 2012b), implying that very small differences are allowed and thus finding a compromise between zero and no constraints, through which both model fit and latent mean comparison can be established. A Bayesian approach involves the use of (1) prior distributions, which represent background knowledge about the parameters of a model, (2) the likelihood function of the data containing the information about the parameters from the data, and (3) a posterior distribution, which contains one's updated knowledge balancing prior knowledge with observed data.

If most of the items show small differences in intercepts, the application of full AMI is recommended, with "small" implying that parameters of substantive interest do not change in a meaningful way if MI does not fully hold (van de Schoot et al., 2013). In most applications, however, the number of non-invariant parameters might be small with respect to the number of actually invariant ones, but this does not prevent an unacceptable model fit. In these cases Muthén and Asparouhov (2013b) and van de Schoot et al. (2013) recommended a two-step procedure in which parameters that are different between groups (and hence are the major sources of misfit) are detected in step 1, for example by using modification indices provided by the ML estimation, and are allowed to be non-invariant to the extent imposed by partial AMI in step 2. A technical description of the statistical features of these models is beyond the scope of this paper (see Muthén and Asparouhov, 2013b; van de Schoot et al., 2013, for a gentle introduction), which is instead to provide a didactic example of how to establish strong measurement invariance using AMI in the particular case of within-couple data.

MATERIALS AND METHODS

SAMPLE

The sample for this study was UK-based and drawn from the first sweep of the Millennium Cohort Study (MCS). The MCS is a longitudinal study drawing its sample from all live births in the United Kingdom over 12 months, in England, Wales, Scotland and Northern Ireland. The first sweep took place in 2003 when the children were aged 9 months, and later follow-ups at the ages of 3, 5, 7, and 11 (University of London, Institute of Education, Centre for Longitudinal Studies, 2012; for details, see Plewis, 2007). In this study we included families which were present at Sweep 1 ($n = 18.552$). If families had twin or triplet births the child coded as cohort member "a" was included in our sample. As we were interested in ratings of both partners within couples, we selected two-parent-figure families, in which both parent figures were present, were of opposite sex, were one generation older than the child, were not blood relatives, and were biological parents (for details, see Malmberg and Flouri, 2011). For the purpose of our demonstration of the AMI procedure we used a sub-sample of parents from the Northern Ireland-advantaged stratum ($n = 527$). Although the procedure presented here can handle missing data either with a full information or a multiple imputation approach, the implications of dealing with missingness were beyond the scope of the paper. As we aimed to provide a relatively straightforward example of the application of BSEM-AMI, we screened for missing data on the GRIMS and used a listwise sample of 435 couples.

MEASURE

The 7-item GRIMS provides a self-reported assessment of the quality of a couple's relationship asking participants to rate seven items [(1) My partner is sensitive to and aware of my needs; (2) My partner doesn't listen to me anymore; (3) I'm sometimes lonely when I'm with my partner; (4) Our relationship is full of joy and excitement; (5) I wish there was more warmth and affection between us; (6) I suspect we are on the brink of separation; (7) We can make up quickly after an argument] on a 5-point, Likert-type agreement scale (1 = Strongly Agree, 5 = Strongly Disagree). Scores of items 1, 4, and 7 were reverse scored before performing the analyses so that higher item scores corresponded to higher RQ.

ANALYTIC STRATEGY

Before testing measurement invariance across partners, we tested the fit of the hypothesized one-factor model for the GRIMS separately for men and women through the "classical" confirmatory factor analysis (CFA) using maximum likelihood estimation with robust standard errors (MLR) to address the relatively non-normal distributions of item scores (Table S1 in Supplementary Information). The fit of the models was evaluated considering the root-mean-square error of approximation (RMSEA), the

Tucker–Lewis index (TLI), and the comparative fit index (CFI), as operationalized in Mplus v7 (Muthén and Muthén, 1998-2012) in association with the MLR estimator. For both the TLI and CFI, values greater than 0.90 and 0.95, respectively, typically reflect acceptable and excellent fit to the data. For the RMSEA, values less than 0.05 and 0.08 reflect a close fit and a reasonable fit to the data, respectively (Marsh et al., 2004). After finding that the expected measurement model fitted adequately in both partners, we specified the sequence of invariance models as in the Meredith (1993) tradition. However, as pointed out by South et al. (2009), testing measurement invariance on paired groups of observations as couple data is different from testing it on independent groups, since both partners are reporting on the same relationship. Instead of testing the same, e.g., one-factor measurement model on two different groups defined by a grouping variable, we therefore modeled the data at the couple level, i.e., the unit of analysis was the couple, and partners were treated as different raters on the same relationship. This means that, for each couple, both women's and men's ratings were on the same line of data. This model is basically equivalent to a single-group two-correlated-factor model in which the items of the scale are considered twice, as indicators of women's and men's marital satisfaction (**Figure 1**).

This also implies that the systematic residual variance (uniqueness) in each pair of identical items between parents is expected to covary because of the identical nature of the item pair (Brown, 2006, chap. 7, e.g., the residual variance in the item "My partner doesn't listen to me any more" for women should covary with the same item for men). Hence, the model with correlated uniqueness (θ_{WiMi}s in **Figure 1**) should have resulted in a substantial improvement in fit over the model without the correlated uniqueness (e.g., Burns et al., 2008). The fit of invariance models was evaluated with the same criteria stated above, while model comparisons were performed using the Satorra-Bentler Scaled Chi-Square Difference Test (Satorra and Bentler, 2001) but, considering that this test suffers the same problems (i.e., sample size dependency) as the chi-square test used to test goodness of fit that led to the development of fit indices, we also considered as support for the more parsimonious model a change in CFI of less than 0.01 or a change in RMSEA of less than 0.015 (Chen, 2007). In case of rejection of the more parsimonious (i.e., constrained) model we inspected modification indices in Mplus output to find the least invariant parameter and re-specified the model letting it to be non-invariant. This procedure was iterated until no more parameters were suggested to be non-invariant.

For Bayesian models we used default prior settings, i.e., normal prior distributions for the intercepts and factor loadings with a prior mean of zero and a prior variance of 10^{10}, and an inverse gamma distribution for the (residual) variance terms with hyperparameters −1 and zero; note that this model is similar to the configural invariance model because it implies practically no "real" prior constraint, and the following Mplus Analysis settings: BCONVERGENCE = 0.01; BITERATIONS = 1,000,000 (20,000); PROCESSOR = 2; CHAINS = 2; BSEED = 167. As indices of model fit we used the posterior predictive *p*-value (PPP) and the 95% confidence interval (CI) for the difference in the *f* statistic for the real and replicated data (see Muthén and Asparouhov, 2012a). An acceptably fitting model should have shown a PPP higher than 0.05 and a 95% CI of the replicated chi-square values that included zero. Nine different wiggle rooms σ^2 (0.50, 0.25, 0.125, 0.05, 0.025, 0.01, 0.005, 0.001, 0.0005) were specified for non-invariant parameters, with smaller values allowing smaller wiggle rooms and therefore a closer approximation of the "classical" invariance model. Given the didactic aim of this paper, we initially tested full AMI models, in which all parameters of interest were allowed to be non-invariant to the extent allowed by the wiggle room, and then turned to partial AMI, following the two-step procedure recommended by Muthén and Asparouhov (2013b) and described above.

The data and all syntax files are available as supplementary materials.

RESULTS

If we simply compared women's and men's observed scores on the GRIMS with a paired-sample *t*-test, we would have concluded that women ($M = 29.23$, $SD = 3.63$, Cronbach's $\alpha =$

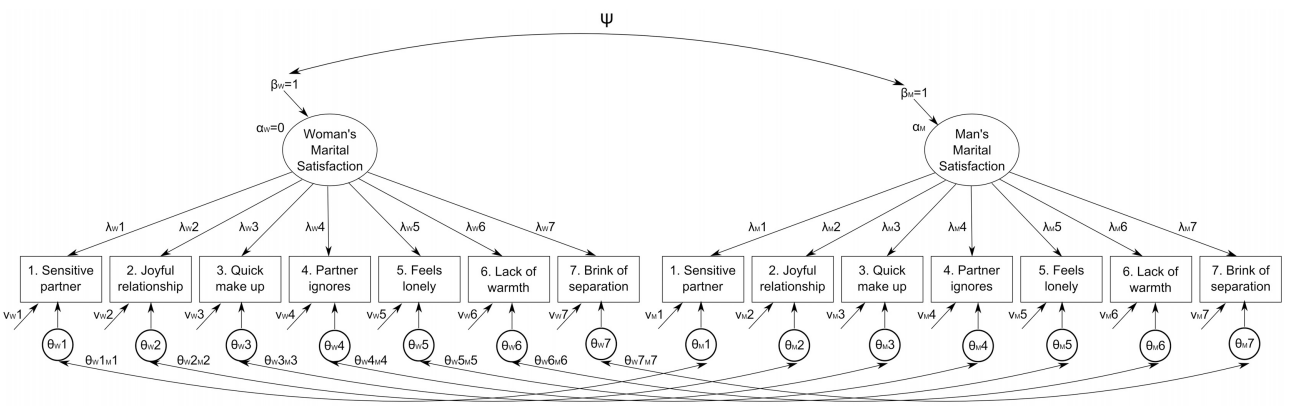


FIGURE 1 | Baseline model used for the confirmatory factor analytic invariance analysis between women's (W subscript) and men's (M subscript) ratings of marital satisfaction. Note that Mplus notation is used, i.e., α, factor mean; β, factor variance; ψ, factor correlation; λ, factor loading; ν, item intercept; θ, item uniqueness.

0.73) tend to be systematically more satisfied than men ($M = 28.19$, $SD = 3.67$, $\alpha = 0.73$) [$t_{(434)} = 5.08$, $p < 0.001$, $d = 0.29$] and that the two scores are only moderately correlated ($r = 0.31$, $p < 0.001$). However, as stated above, this result is meaningful only if strong measurement invariance holds. The one-factor measurement model for the GRIMS had an acceptable fit for women [$SB\chi^2(14) = 33.258$, $p = 0.002$, Scaling Correction Factor [SCF] = 1.071, CFI = 0.944, TLI = 0.916, RMSEA = 0.056) and optimal for men [$SB\chi^2(14) = 16.862$, $p = 0.264$, SCF = 1.491, CFI = 0.990, TLI = 0.985, RMSEA = 0.022]. Factor score determinacies, i.e., validity coefficients computed as the correlation between factor score estimates and their respective latent factors, were 0.868 and 0.871, respectively, suggesting a high (>0.80, Gorsuch, 1983) degree of convergence of observed scores on the scale and the latent individuals' scores. Raykov (1997)'s composite reliabilities were in both cases 0.740.

As a first step for testing measurement invariance we compared the configural invariance models with and without correlated uniquenesses. As shown in **Table 1**, the fit of both models was acceptable, suggesting an adequate ability of the a priori one-factor measurement model to fit the data in each partner without invariance constraints. However, the model with the correlated uniqueness fitted statistically and substantially better than the one without, and was thus chosen as the baseline model for the invariance tests.

We then constrained the loadings for identical items to be equal between parents (*weak* or *metric invariance*). If identical items have statistically equivalent loadings, then the identical items show the same amount of increase between parents for the same amount of increase on the latent factor. As shown in **Table 1**, this constraint did not significantly affected model fit, hence we concluded that weak invariance held. However, the comparison of latent means is appropriate only if it can be shown that also intercepts of the same items are invariant between partners, i.e., *strong* or *scalar invariance* holds. When factor loadings and intercepts are invariant, at any point along the factor continuum the same level of the factor results in statistically equivalent average scores on identical items between parents, namely, any observed score differences between parents on identical items is not due to partner bias but rather to actual differences on the factor mean. Due to model identification issues it is not possible to estimate the two latent means simultaneously, hence we fixed at zero the women's mean and estimated the men's mean, which thus represented the mean difference of men's RQ scores with respect to women's. **Table 1** shows that constraining to invariance all item intercepts led to a substantial and significant decrease of fit, although this remained acceptable. We concluded that full strong invariance could not be assumed, thus undermining the possibility of reliably comparing latent means.

The inspection of modification indices from the strong invariance model indicated that intercept of item 4 should be allowed to vary across parents. The fit of the model allowing for the partial invariance of this intercept (Partial1 in **Table 1**) was higher than the fit of the strong invariance model, but it was still lower than that of the weak invariance model [$SB\Delta\chi^2(5) = 18.692$, $p = 0.002$ and $\Delta CFI > 0.01$]. Following the modification indices, in a subsequent model (Partial2) we permitted the intercept of item 3

Table 1 | Goodness of fit of maximum likelihood with robust standard errors (MLR) full and partial invariance models.

Invariance model	$SB\chi^2$	*df*	*p*	SCF	RMSEA	CFI	TLI	*cd*	TRd	Δdf	*p*	Mean difference estimate				Simulation					Bias	
												α	*SE*	*p*	*d*	AVG	*SD*	*SE* avg	95% Cover	% Sig.	*M* bias (%)	*SE* Bias (%)
Configural no CUs	97.948	76	0.046	1.112	0.026	0.973	0.967															
Configural with CUs	82.052	69	0.135	1.111	0.021	0.984	0.979	1.112	15.889	7	0.026											
Weak with CUs	90.371	76	0.125	1.115	0.021	0.982	0.979	1.154	8.319	7	0.305											
Strong with CUs	127.238	82	0.001	1.108	0.036	0.944	0.937	1.017	39.534	6	<0.001	−0.341	0.070	<0.001	0.25	−0.344	0.072	0.068	0.938	0.998	0.79	−5.58
Partial1^ with CUs	108.190	81	0.024	1.110	0.028	0.966	0.962	0.946	22.081	1	<0.001											
Partial2* with CUs	94.677	80	0.126	1.110	0.021	0.982	0.979	1.110	13.513	1	<0.001	−0.502	0.075	<0.001	0.35	−0.505	0.079	0.075	0.945	1.000	0.56	−4.20

*$SB\chi^2$, Satorra-Bentler scaled chi-square; df, degrees of freedom; SCF, Scaling Correction Factor; RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index; TLI, Tucker-Lewis Index; cd, difference test scaling correction; TRd, Satorra-Bentler scaled chi-square difference test; Δdf, degrees of freedom difference; CU, correlated uniquenesses; ^Intercept of item 4 was not invariant; *Intercepts of items 3 and 4 were not invariant. SB scaled chi-square difference tests are referred to the model in the above line.*

be invariant, too. This model fitted significantly and substantially better than Partial1 model and its fit did not differ from the fit of the weak invariance model [SB$\Delta\chi^2(5) = 4.264$, $p = 0.371$ and ΔCFI < 0.01]. The standardized estimated latent mean difference was −0.502, and was statistically different from zero, suggesting that women were more satisfied with their relationship than men, although with a small ($d < 0.50$) effect size.

We then turned to Bayesian SEM, and re-analyzed the configural, weak and strong invariance models (FMI1–FMI3 in **Table 2**).

As shown in **Table 2**, the "classical" strong invariance model (all intercepts constrained to equality across partners, FMI3), did not fit the data, since the posterior predictive *p*-value was <0.05, and the 95% CI of the replicated chi-square values did not include zero, whereas the configural (FMI1) and weak (FMI2) invariance model adequately fitted the data, but did not allow to compare the latent means. We thus resorted to full AMI and we restricted intercept differences by specifying the 9 prior distributions described above (same "wiggle room" for all intercepts, AFMI1–AFMI9 in **Table 2**). Results are shown in the upper part of **Table 2**, and values in the median absolute intercept difference column show that restricting the wiggle room led to smaller intercept differences. Models with prior variance 0.001 (AFMI8) and 0.0005 (AFMI8) should be rejected, since either their 95% CI for the difference between the observed and the replicated χ^2 did not include zero or their ppp-value was lower than 0.05, or both. However, it is interesting to note that, among the acceptably fitting models, the estimate of the factor mean difference was not always significantly different from zero, probably due to alignment (see Discussion). It became so only when σ^2 was 0.05 (AFMI4) or lower, with small effect sizes.

The Mplus output for Bayesian AMI models provides the equivalent of modification indices in ML MI models, i.e., the DIFFERENCE OUTPUT, in which the deviations from the mean and their significance for non-invariant parameters are shown. Deviations from the mean were significant for item 4 in model AMI5, for items 4 and 5 in model AMI6, for items 3, 4, and 5 in model AMI7 and for item 4 in model AMI8 (Table S2 in Supplementary Materials). In order to compare the results with the ML partial invariance models, we tested approximate partial measurement invariance (APMI) models allowing a wiggle room only for items 3 and 4 (PMI1 and APMI1–APMI9 in **Table 2**), while constraining to equality the intercepts of the other items. Note that instead of the modification index approach of relaxing one equality restriction at a time, we followed Muthén and Asparouhov (2013b)'s suggestion to relax all misfitting equalities, since when they are not too many they do not have much effect on the point estimates nor on the identification of the model, although they might slightly increase the standard errors. As it is shown in the bottom part of **Table 2**, an adequate fit was obtained when σ^2 ranged from 0.50 to 0.005 (APMI1–APMI6), and, in these cases, all estimates of factor mean differences were statistically significant. Effect sizes were larger than in full invariance models and similar to the effect size of the partial invariance ML model, but still in the small range.

Given the mixed pattern of results about the estimate of latent mean differences, we wondered which result should be trusted. Hence, we investigated the possible bias in the comparison of latent means through a Monte Carlo simulation study. van de Schoot et al. (2013) investigated the possible bias in the comparison of latent means as a result of applying the approximate MI model by performing a simulation study in which seven populations with different sets of (assumed) true values were specified. Since we could not know the true population values, we decided to use the estimates obtained in testing the model as population values to explore the stability of the models and the appropriate convergence of parameter estimates to the assumed population parameters. Results were obtained with ESTIMATOR = ML and with ESTIMATOR = BAYES. For the latter we used PROCESSORS = 2; BCONVERGENCE = 0.01; BITERATIONS = (5000); BSEED = 167; and the default priors for both full and partial invariance models.

For each population we generated 1000 datasets. We considered an estimate as acceptably unbiased if (1) the empirical standard deviation of the 1000 estimated mean differences was lower than 0.10; (2), the relative mean bias of the estimate defined as (AVG-α)/α)*100, where AVG is average mean obtained from the simulation study and α is the assumed population value, did not exceed 10% (e.g., van de Schoot et al., 2013); (3) the standard error bias for the parameter for the mean difference parameter did not exceed 5% (Muthén and Muthén, 2002); (4) 95% coverage, i.e., the proportion of replications for which the 95% CI included the population value, was at least 95%; (5), the significance criterion, i.e., the proportion of datasets for which the 95% CI of the factor mean difference estimate did not include zero and was therefore statistically significant, was close to 1.00. Results are shown in the rightmost columns of **Table 2**.

Among the adequately fitting full invariance models, no full invariance model met the criteria stated above, since the standard error bias was always higher than 5%. The simulation of the full strong invariance model with ML did not meet criteria 3 and 4 (**Table 1**). On the other hand, PMI and APMI models 1–5 appeared to provide sufficiently unbiased results, although the standard error bias was slightly over the cut-off (**Table 2**). The simulation of the partial strong invariance model with ML basically met all criteria, with only criterion 4 borderline met (**Table 1**).

Taken together, these results suggest that there are gender differences in the perception of RQ as measured by the 7-item GRIMS in intact couples, with women reporting higher scores.

It is interesting to note that estimates and significance of the factor correlation were stable at slightly less than 0.42 throughout all models, suggesting that a higher relationship satisfaction in women tends to be associated with a higher relationship satisfaction in men, and that all the criteria stated above for assessing bias were basically satisfied (only criterion 4 showed borderline values; see Table S3 in the Supplementary Materials).

DISCUSSION

The aim of this study was to demonstrate the usefulness and flexibility of the BSEM-AMI for investigating measurement invariance, particularly addressing the issue of possible lack of strong invariance in within-couple data. Specifically, we provided an example of how a common issue in relationship research, i.e., partner differences in the perception of relationship quality, can

Table 2 | Goodness of fit, estimated latent mean differences and their estimated bias for Bayesian full and partial invariance models.

Model	95% CI χ^2	PPP	Median absolute intercept difference (range)	Mean difference estimate				Simulation					Bias	
				α	*SE*	*p*	*d*	AVG	*SD*	*SE* Avg	95% Cover	% Sig.	*M* Bias (%)	*SE* Bias (%)
FULL INVARIANCE														
FMI1 configural	−16.183 58.471	0.106		–	–	–		–	–		–	–	–	
FMI2 weak	−12.794 57.036	0.111	0.131 (0.010–0.409)	–	–	–		–	–		–	–	–	
FMI3 strong	20.836 93.107	0.001	0.000 (0.000–0.000)	−0.340	0.066	<0.001	0.27	−0.348	3.161	0.593	0.314	0.682	2.24	−81.23
AFMI1 $\sigma^2 = 0.5$	−14.605 59.226	0.119	0.088 (0.014–0.224)	−0.279	0.533	0.284	0.03	−0.340	0.360	0.425	0.952	0.123	22.01	18.01
AFMI2 $\sigma^2 = 0.25$	−14.219 58.635	0.119	0.071 (0.007–0.196)	−0.315	0.377	0.194	0.04	−0.344	0.194	0.346	0.995	0.080	9.24	78.25
AFMI3 $\sigma^2 = 0.125$	−14.811 59.127	0.122	0.062 (0.001–0.184)	−0.333	0.280	0.123	0.06	−0.340	0.117	0.266	1.000	0.119	1.98	127.76
AFMI4 $\sigma^2 = 0.05$	−15.240 61.016	0.107	0.051 (0.003–0.183)	−0.354	0.185	0.027	0.10	−0.342	0.077	0.186	1.000	0.410	−3.50	142.00
AFMI5 $\sigma^2 = 0.025$	−15.163 61.056	0.108	0.048 (0.004–0.178)	−0.355	0.138	0.005	0.13	−0.342	0.069	0.142	1.000	0.827	−3.80	104.18
AFMI6 $\sigma^2 = 0.01$	−14.855 61.681	0.106	0.044 (0.003–0.158)	−0.350	0.101	<0.001	0.18	−0.340	0.066	0.104	0.997	0.976	−2.94	56.97
AFMI7 $\sigma^2 = 0.005$	−13.530 63.776	0.086	0.036 (0.001–0.134)	−0.346	0.084	<0.001	0.21	−0.338	0.065	0.087	0.987	0.995	−2.43	34.41
AFMI8 $\sigma^2 = 0.001$	−0.366 77.232	0.028	0.016 (0.000–0.062)	−0.341	0.069	<0.001	0.26	−0.333	0.064	0.071	0.967	0.999	−2.38	9.97
AFMI9 $\sigma^2 = 0.0005$	6.400 83.511	0.012	0.009 (0.000–0.038)	−0.341	0.067	<0.001	0.26	−0.333	0.064	0.068	0.961	1.000	−2.29	6.88
PARTIAL INVARIANCE*														
PMI strong	−13.704 62.416	0.095	−0.214 −0.264	−0.499	0.071	<0.001	0.36	−0.494	0.070	0.075	0.967	1.000	−1.10	6.28
APMI1 $\sigma^2 = 0.5$	−11.802 61.498	0.090	−0.213 −0.263	−0.499	0.071	<0.001	0.36	−0.492	0.071	0.074	0.967	1.000	−1.34	5.09
APMI2 $\sigma^2 = 0.25$	−11.718 61.594	0.090	−0.212 −0.262	−0.497	0.074	<0.001	0.35	−0.489	0.071	0.074	0.965	1.000	−1.59	5.25
APMI3 $\sigma^2 = 0.125$	−11.799 61.592	0.090	−0.209 −0.259	−0.495	0.074	<0.001	0.35	−0.485	0.070	0.074	0.964	1.000	−2.12	5.12
APMI4 $\sigma^2 = 0.05$	−12.175 61.725	0.087	−0.199 −0.250	−0.489	0.074	<0.001	0.34	−0.472	0.069	0.073	0.959	1.000	−3.46	6.24
APMI5 $\sigma^2 = 0.025$	−11.671 62.672	0.087	−0.186 −0.236	−0.479	0.074	<0.001	0.33	−0.454	0.068	0.072	0.954	1.000	−5.14	6.33
APMI6 $\sigma^2 = 0.01$	−9.953 64.009	0.081	−0.155 −0.203	−0.457	0.072	<0.001	0.33	−0.419	0.066	0.071	0.934	1.000	−8.29	6.49
APMI7 $\sigma^2 = 0.005$	−7.646 67.449	0.062	−0.122 −0.165	−0.433	0.071	<0.001	0.32	−0.388	0.065	0.069	0.921	1.000	−10.39	6.45
APMI8 $\sigma^2 = 0.001$	6.166 81.574	0.017	−0.045 −0.067	−0.374	0.069	<0.001	0.28	−0.340	0.064	0.067	0.933	0.999	−9.14	4.35
APMI9 $\sigma^2 = 0.0005$	12.325 87.639	0.008	−0.025 −0.039	−0.358	0.068	<0.001	0.27	−0.334	0.064	0.067	0.945	1.000	−6.70	4.07

*95% CI χ^2, 95% confidence interval for the difference between the observed and the replicated χ^2; PPP, posterior predictive p-value; α, mean latent score difference parameter; SE, standard error of α; p, significance of α; d, effect size; AVG, average of estimated mean latent score differences over the replications; SD, standard deviation of mean difference parameter estimate over the replications; SE avg, average of the estimated standard errors for the mean difference parameter estimate over the replications; 95% cover, proportion of replications for which the 95% CI included the hypothesized population value α; % sig., proportion of datasets for which the 95% CI did not include zero, i.e., the percentage of datasets for which it can be concluded that AVG is larger than zero in the population; M bias, (AVG-α)/α *100; SE bias, (SE avg-SD)/SD; FMI, Full measurement invariance; AFMI, Approximate full measurement invariance; PMI, partial measurement invariance; APMI, approximate partial measurement invariance; σ^2, prior variance of intercepts for all pairs of items; *Intercepts of items 3 and 4 were not invariant, values in the intercept difference column are the differences between women's–men's intercepts for items 3 and 4, respectively.*

be addressed with this methodology. We applied BSEM-AMI to ratings on the 7-item GRIMS in a sample of intact couples drawn from the MCS database, and the results suggested that women perceived a higher RQ than men (although with a small effect size), somewhat contradicting the results of a recent meta-analysis (Jackson et al., 2014) that showed that in intact couples there are no substantial differences in RQ. However, this meta-analysis did not include any study using the 7-item GRIMS, nor could we screen our sample for couples in marital therapy, for which partner differences in RQ are known to exist, though in the other direction (Jackson et al., 2014). As a limitation to the generalizability of the results for the substantive issue of this paper, it must also be considered that the data used in this study were collected in a specific subgroup of couples, i.e., Northern-Irish, advantaged, intact couples 9 months after the birth of a child. This sample is similar to South et al. (2009)'s study, in which data from intact couples with long-term marriages included in the Minnesota Twin Family Study were used and results similar to ours were found, as women reported higher levels of dyadic adjustment. These results are also consistent with those of Shapiro et al. (2000), who reported that marital satisfaction was significantly higher for women who became mothers than for men who became fathers. However, since other studies suggested that marital satisfaction is lower among the individuals who are most responsible for the child, which in most cases is the mother (e.g., Hochschild, 1989), and in light of the aforementioned limitations, we cannot consider our results conclusive as to the general question of whether women and men differ in the perception of RQ.

In pursuing the substantive focus of this investigation, we showed how BSEM-AMI can be successfully applied to address it. As we found lack of support for a strong invariance model, i.e., a model that assumes that all item intercepts, along with factor loadings, are perfectly invariant across partners and allows a valid comparison of latent scores, through BSEM-AMI we could release the assumption of a *zero* difference between intercepts and allow a "wiggle-room" for it, i.e., an *approximately* zero difference. AMI models that allowed this wiggle room for all intercepts showed an acceptable fit and suggested that differences in GRIMS scores could exist between partners. However, as shown in **Table 2**, the significance of the mean difference parameter α increased as the wiggle room got smaller. This result might be due to alignment, i.e., a parameterization indeterminacy (Muthén and Asparouhov, 2013b). In other words, the BSEM-AMI tries to find a solution in which the variance across partners for a measurement parameter is small. Since the wiggle room is prior variance distribution, and thus allows a pre-determined range of variation for parameter estimates, the method is more effective when there is a large degree of minor non-invariance and parameters deviations from invariance are in opposite directions and can largely cancel each other out (Muthén and Asparouhov, 2013b). In specifying full AMI models, we assumed that the wiggle room was the same for all intercepts: in these cases, however, if there is an item with a relatively large difference whereas the difference is relatively small in all the others, the BSEM small-variance prior for the parameter differences tends to pull the deviating parameter toward the average of the parameters for both partners. This means that the deviating parameter will be smaller and the invariant parameters larger than their true values. With intercepts misestimated, the factor means and factor variances are misestimated too (Muthén and Asparouhov, 2013b). Our simulation studies suggested that this might have been the case of the present investigation (due to the relatively large difference in intercepts for items 3 and 4), given the BSEM analysis we used is not expected to always recover parameter values used to simulate the data. Although the mean difference estimates showed a negligible bias, the standard error bias was large. Muthén and Asparouhov (2013a) noted that this does not necessarily mean that the model does not fit the data, but that an equally well-fitting solution with a possibly simpler interpretation due to another simplicity criterion may be available. They also suggest to detect non-invariant items and relax equalities only on them, since this will lead to the recovering of the parameter values. When we resorted to partial invariance models, and we identified in intercepts of items 3 and 4 the major sources of lack of fit, results were much more stable and unbiased, although it seemed that a very small wiggle room still lead to inadequately fitting models (see AMPI models 8–9 in **Table 2**). Actually, the ML partial invariance model also provided a good fit and unbiased estimated of mean difference and its standard error, suggesting that, in this case, a more "classical" approach would have led to similar conclusions as the "new" Bayesian approach.

The results of this study seem to support the efficacy of the two-step approach suggested by Muthén and Asparouhov (2013b) and van de Schoot et al. (2013), in which parameters that are different between groups are detected in step 1 through modification indices and are allowed to be non- (or approximately) invariant in step 2. However, it should be noted that partial measurement invariance has the main shortcoming that the parameters to be freed are identified with an ex-post facto procedure. This might raise the issue of capitalization on chance when the sample is small and/or not representative of the population, and undermine the generalizability of the results, especially in those studies in which the measurement model of a scale is investigated. Muthén and Asparouhov (2012a) pointed out that while ML modification indices inform about model improvement when a single parameter is freed and can lead to a long series of modifications, BSEM can inform about model modification when all parameters are freed and does so in a single step: their simulations showed sufficient power to detect model misspecification in terms of 95% Bayesian credibility intervals not covering zero. Nontheless, they also warn that, as with ML model modification, BSEM model modification should be supported by substantive interpretability. However, in most research contexts (e.g., developing a new questionnaire) one cannot know in advance the sources of non-invariance and whether deviations from invariance will eventually bias the substantive conclusions or not: hence we recommend to carefully check the stability of model estimates through simulation studies.

Another issue that it is worth noting, although not a focus of this study, is that the indices of model fit for Bayesian and ML SEM do not always overlap. The reason for which we chose the Northern Ireland-advantaged subsample of the whole MCS dataset is that it was the only one in which the fit indices of

the one-factor measurement model for the GRIMS were adequate for both estimation methods. When we considered the largest stratum of the MCS, i.e., the England advantaged stratum ($n = 3830$), the configural invariance model fitted adequately with ML, but it was not even remotely adequate with BSEM (see Table S4 in Supplementary Materials). As an example, we drew random smaller subsamples and retested model fit, finding no substantial changes in ML indices, but a gradual approach to acceptable values for BSEM. Muthén and Asparouhov (2012a) warn that the posterior PPP does not behave like a *p*-value for a chi-square test of model fit (e.g., Hjort et al., 2006), hence the Type I error is not 5% for a correct model. Since there is not a theory for how low the PPP can be before the model is significantly inadequately fitting at a certain level, Muthén and Asparouhov (2012a) consider it more akin to an SEM fit index rather than a chi-square test. Using simulations they found that PPP performed better than the ML likelihood-ratio chi-square test at small sample sizes where ML typically inflates chi-square and that it was less sensitive than ML to ignorable deviations from the correct model thus concluding that PPP seems to have sufficient power to detect important model misspecifications, that might go unnoticed when using ML.

ACKNOWLEDGMENT

We are grateful to The Centre for Longitudinal Studies, Institute of Education for the use of these data and to the UK Data Archive and Economic and Social Data Service for making them available. However, they bear no responsibility for the analysis or interpretation of these data.

We also thank Prof. John Rust for kindly sharing with us information about the development of the 7-item GRIMS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00983/abstract

REFERENCES

Bernard, J. (1972). *The Future of Marriage*. New Haven, CT: Yale University Press.

Bradbury, T. N., Fincham, F. D., and Beach, S. R. H. (2000). Research on the nature and determinants of marital satisfaction: a decade in review. *J. Marriage Fam.* 62, 964–980. doi: 10.1111/j.1741-3737.2000.00964.x

Bronte-Tinkew, J., Guzman, L., Jekielek, S., Moore, K. A., Ryan, S., Redd, Z., et al. (2004). *Conceptualizing and Measuring "Healthy Marriage" for Empirical Research and Evaluation Studies: A Review of the Literature and Annotated Bibliography.* Washington, DC: Child Trends, Inc. Available online at: http://www.childtrends.org/wp-content/uploads/2014/03/2003-24CHealthyMarriageLitReviewBibliographyExecSum.pdf

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Press.

Burns, G. L., de Moura, M. A., Walsh, J. A., Desmul, C., Silpakit, C., and Sommers-Flanagan, J. (2008). Invariance and convergent and discriminant validity between mothers' and fathers' ratings of oppositional defiant disorder toward adults, ADHD–HI, ADHD–IN, and academic competence factors within Brazilian, Thai, and American Children. *Psychol. Assess.* 20, 121–130. doi: 10.1037/1040-3590.20.2.121

Busby, D. M., and Christensen, C. (1995). A revision of the dyadic adjustment scale for use with distressed and nondistressed couples: construct hierarchy and multidimensional scales. *J. Marital Fam. Ther.* 21, 289–308. doi: 10.1111/j.1752-0606.1995.tb00163.x

Byrne, B. M., Shavelson, R. J., and Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Model.* 14, 464–504. doi: 10.1080/10705510701301834

Cheung, G. W., and Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equation modeling. *J. Cross Cult. Psychol.* 31, 187–212. doi: 10.1177/0022022100031002003

Child Trends. (2003). *Conceptualizing and Measuring "Healthy Marriages" for Empirical Research and Evaluation Studies: A Compendium of Measures- Part II.* Washington, DC: Child Trends, Inc. Available online at: http://www.childtrends.org/wp-content/uploads/2013/09/Healthy-Marriages-Part-II.pdf

Fincham, F. D. (2003). Marital conflict: correlates, structure, and context. *Curr. Dir. Psychol. Sci.* 12, 23–27. doi: 10.1111/1467-8721.01215

Gorsuch, R. L. (1983). *Factor Analysis, 2nd Edn*. Hillsdale, NJ: Erlbaum.

Grych, J. H., and Fincham, F. D. (eds.). (2001). *Interparental Conflict and Child Development: Theory, Research, and Applications.* New York, NY: Cambridge University Press.

Hjort, N. L., Dahl, F. A., and Steinbakk, G. H. (2006). Post-processing posterior predictive p-values. *J. Am. Stat. Assoc.* 101, 1157–1174. doi: 10.1198/016214505000001393

Hochschild, A. R. (1989). *The Second Shift*. New York, NY: Avon Books.

Jackson, J. B., Miller, R. B., Oka, M., and Henry, R. G. (2014). Gender differences in marital satisfaction: a meta-analysis. *J. Marriage Fam.* 76, 105–129. doi: 10.1111/jomf.12077

Jones, E. (2010). "Parental relationship and parenting," in *Children of the 21st Century. The First Five Years,* eds K. Hansen, H. Joshi and S. Dex (Bristol: The Policy Press), 53–76.

Kenny, D. A. (1996). Models of non-independence in dyadic research. *J. Soc. Pers. Relat.* 13, 279–294. doi: 10.1177/0265407596132007

Kenny, D. A., and Cook, W. (1999). Partner effects in relationship research: conceptual issues, analytic difficulties, and illustrations. *Pers. Relat.* 6, 433–448. doi: 10.1111/j.1475-6811.1999.tb00202.x

Locke, H. J., and Wallace, K. M. (1959). Short marital adjustment and prediction tests: their reliability and validity. *Marriage Fam. Living* 21, 251–255. doi: 10.2307/348022

Luo, S., Chen, H., Yue, G., Zhang, G., Zhaoyang, R., and Xu, D. (2008). Predicting marital satisfaction from self, partner, and couple characteristics: is it me, you, or us? *J. Pers.* 76, 1231–1265. doi: 10.1111/j.1467-6494.2008.00520.x

Malmberg, L.-E., and Flouri, E. (2011). The comparison and interdependence of maternal and paternal influences on child mental health and resilience. *J. Clin. Child Adolesc.* 40, 1–11. doi: 10.1080/15374416.2011.563469

Marsh, H. W., Hau, K.-T., and Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Struct. Equ. Model.* 11, 320–341. doi: 10.1207/s15328007sem1103_2

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Muthén, B., and Asparouhov, T. (2012a). Bayesian SEM: a more flexible representation of substantive theory. *Psychol. Methods* 17, 313–335. doi: 10.1037/a0026802

Muthén, B., and Asparouhov, T. (2012b). *New Features in Mplus v7 Lecture 3.* Available online at: http://mplus.fss.uu.nl/2012/09/12/the-workshop-new-features-of-mplus-v7/

Muthén, B., and Asparouhov, T. (2013a). "Item response modeling in Mplus: a multi-dimensional, multi-level, and multi-timepoint example," in *Handbook of Item Response Theory: Models, Statistical Tools, and Applications*, eds W. J. van der Linden and R. K. Hambleton (Boca Raton, FL: Chapman & Hall/CRC Press). Available online at: http://www.statmodel.com/download/IRT1Version2.pdf

Muthén, B., and Asparouhov, T. (2013b). *BSEM Measurement Invariance Analysis. Mplus WebNotes: No. 17.* Available online at: www.statmodel.com

Muthén, B. O., and Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika* 46, 407–419. doi: 10.1007/BF02293798

Muthén, L. K., and Muthén, B. O. (1998-2012). *Mplus User's Guide, 7th Edn*. Los Angeles, CA: Muthén and Muthén. Available online at: www.statmodel.com

Muthén, L. K., and Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Struct. Equ. Model.* 9, 599–620. doi: 10.1207/S15328007SEM0904_8

Plewis, I. (ed.). (2007). *The Millennium Cohort Study Technical Report on Sampling, 4th Edn*. Available online at: http://www.cls.ioe.ac.uk/library-media%5Cdocuments%5CTechnical_Report_on_Sampling_4th_Edition.pdf

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Appl. Psychol. Meas.* 21, 173–184. doi: 10.1177/014662169702 12006

Reynolds, J., Houlston, C., and Coleman, L. (2014). *Understanding Relationship Quality*. London: OnePlusOne. Available online at: http://www.oneplusone.org.uk/wp-content/uploads/2014/02/UnderstandingRelationship-Quality-by-Jenny-Reynolds-Dr-Catherine-Houlston-and-Dr-Lester-Coleman.pdf

Robles, T. F., Slatcher, R. B., Trombello, J. M., and McGinn, M. M. (2014). Marital quality and health: a meta-analytic review. *Psychol. Bull.* 140, 140–87. doi: 10.1037/a0031859

Rust, J., Bennun, I., Crowe, M., and Golombok, S. (1986). The Golombok Rust inventory of marital state (GRIMS). *Sex. Marital Ther.* 1, 55–60. doi: 10.1080/02674658608407680

Rust, J., Bennun, I., Crowe, M., and Golombok, S. (1990). The GRIMS. A psychometric instrument for the assessment of marital discord. *J. Fam. Ther.* 12, 45–57. doi: 10.1046/j.1990.00369.x

Rust, J., and Golombok, S. (1985). The Golombok-Rust inventory of sexual satisfaction (GRISS). *Br. J. Clin. Psychol.* 24, 63–64. doi: 10.1111/j.2044-8260.1985.tb01314.x

Satorra, A., and Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika* 66, 507–514. doi: 10.1007/BF02296192

Shapiro, A. F., Gottman, J. M., and Carrère, S. (2000). The baby and the marriage: identifying factors that buffer against decline in marital satisfaction after the first baby arrives. *J. Fam. Psychol.* 14, 59–70. doi: 10.1037/0893-3200.14.1.59

South, S. C., Krueger, R. F., and Iacono, W. G. (2009). Factorial invariance of the Dyadic Adjustment Scale across gender. *Psychol. Assess.* 21, 622–628. doi: 10.1037/a0017572

Spanier, G. B. (1976). Measuring dyadic adjustment: new scales for assessing the quality of marriage and similar dyads. *J. Marriage Fam.* 38, 15–28. doi: 10.2307/350547

University of London, Institute of Education, Centre for Longitudinal Studies (2012). *Millennium Cohort Study: First Survey, 2001-2003 [Computer File], 11th Edn*. Colchester: UK Data Archive [distributor]. doi: 10.5255/UKDA-SN-4683-3

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *Eur. J. Dev. Psychol.* 9, 486–492. doi: 10.1080/17405629.2012.686740

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 May 2014; accepted: 19 August 2014; published online: 19 September 2014.

*Citation: Chiorri C, Day T and Malmberg L-E (2014) An approximate measurement invariance approach to within-couple relationship quality. Front. Psychol. **5**:983. doi: 10.3389/fpsyg.2014.00983*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 08 September 2014
doi: 10.3389/fpsyg.2014.00982

Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values

***Jan Cieciuch*[1,2]*, *Eldad Davidov*[3], *Peter Schmidt*[4,5], *René Algesheimer*[6] and *Shalom H. Schwartz*[4,7]**

[1] University Research Priority Program 'Social Networks,' University of Zürich, Zürich, Switzerland
[2] Institute of Psychology, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland
[3] Institute of Sociology, University of Zürich, Zürich, Switzerland
[4] International Laboratory for Socio-Cultural Research, National Research University-Higher School of Economics, Moscow, Russia
[5] Department of Political Science, University of Giessen, Giessen, Germany
[6] Department of Business Administration, University of Zürich, Zürich, Switzerland
[7] Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, Israel

Edited by:
Rens Van de Schoot, Utrecht University, Netherlands

Reviewed by:
Daniel Saverio John Costa, University of Sydney, Australia
Ben Kelcey, University of Cincinnati, USA
Sebastian Jilke, Erasmus University Rotterdam, Netherlands

****Correspondence:***
Jan Cieciuch, University Research Priority Program 'Social Networks', Andreasstrasse 15, CH-8050 Zürich, Switzerland
e-mail: jancieciuch@gmail.com

One of the most frequently used procedures for measurement invariance testing is the multigroup confirmatory factor analysis (MGCFA). Muthén and Asparouhov recently proposed a new approach to test for approximate rather than exact measurement invariance using Bayesian MGCFA. Approximate measurement invariance permits small differences between parameters otherwise constrained to be equal in the classical exact approach. However, extant knowledge about how results of approximate measurement invariance tests compare to the results of the exact measurement invariance test is missing. We address this gap by comparing the results of exact and approximate cross-country measurement invariance tests of a revised scale to measure human values. Several studies that measured basic human values with the Portrait Values Questionnaire (PVQ) reported problems of measurement noninvariance (especially scalar noninvariance) across countries. Recently Schwartz et al. proposed a refined value theory and an instrument (PVQ-5X) to measure 19 more narrowly defined values. Cieciuch et al. tested its measurement invariance properties across eight countries and established exact scalar measurement invariance for 10 of the 19 values. The current study applied the approximate measurement invariance procedure on the same data and established approximate scalar measurement invariance even for all 19 values. Thus, the first conclusion is that the approximate approach provides more encouraging results for the usefulness of the scale for cross-cultural research, although this finding needs to be generalized and validated in future research using population data. The second conclusion is that the approximate measurement invariance is more likely than the exact approach to establish measurement invariance, although further simulation studies are needed to determine more precise recommendations about how large the permissible variance of the priors may be.

Keywords: multigroup confirmatory factor analysis, exact measurement invariance, approximate measurement invariance, configural metric scalar measurement invariance, revised Portrait Values Questionnaire, Bayesian analysis

MEASUREMENT INVARIANCE

Measurement invariance is a psychometric property of a scale developed to measure a latent construct. The instrument is measurement invariant when the same construct is measured in the same way across different groups, such as countries, cultural units, time points, or regions within countries (Horn and McArdle, 1992; Meredith, 1993; Vandenberg and Lance, 2000; Vandenberg, 2002; Millsap, 2011; Davidov et al., 2014). Measurement invariance is necessary for conducting meaningful comparisons across groups. The most widely used method to establish measurement invariance is multigroup confirmatory factor analysis (MGCFA; Jöreskog, 1971; Bollen, 1989). Usually one distinguishes between three levels of measurement invariance: configural (where all groups have the same pattern of factor loadings), metric (where the factor loadings are constrained to be equal across the compared groups), and scalar (where the factor loadings and the indicator intercepts are constrained to be equal across groups) (Vandenberg and Lance, 2000). Metric invariance is sufficient for comparing covariances and unstandardized regression coefficients across groups. A meaningful comparison of latent means across groups, however, requires the scalar level of measurement invariance.

Some researchers have argued that partial (metric or scalar) measurement invariance is sufficient for meaningful comparisons

(Byrne et al., 1989; Steenkamp and Baumgartner, 1998). Partial invariance is supported when the parameters of at least two indicators (loadings at the metric level and loadings plus intercepts at the scalar level of the measurement) are equal across groups.

Measurement invariance is becoming an increasingly important and disputed topic in the social sciences. To illustrate, in April 2014, the term "measurement invariance" yielded about 239,000 hits in a Google Scholar search. This abundance of scientific papers falls into three categories. The first category includes methodological papers that introduce, discuss, and evaluate various methods and approaches to measurement invariance. The second includes papers that test the measurement invariance of a given construct across groups as a precondition for further comparative analysis. These papers assess measurement invariance as a preliminary analysis that allows for a meaningful test of the substantive hypotheses. The third category of papers reports the measurement invariance properties of specific questionnaires that were developed to measure specific latent constructs. These papers assess the quality of the questionnaires for analyses within and across countries or time points. They seek to improve questionnaire validity and reliability by identifying weaknesses and problems in the formulation of questions, in translation, in culture appropriateness, and so on. Establishing measurement invariance in one study does not signify that a questionnaire is always measurement invariant. Measurement invariance should be repeatedly tested across groups, because noninvariance can be caused by external features of the study in addition to internal features of the instrument.

The aim of the present study is two-fold. First, we try to establish the measurement invariance properties of Schwartz et al.'s (2012) newly developed scale to measure human values. This goal locates the present study in the third category of studies listed above. Second, we apply two methods (exact and approximate) for establishing measurement invariance and compare their findings. This goal locates the present study in the first category of studies listed above. The approximate approach for testing measurement invariance is more liberal than the exact approach. However, extant knowledge about how results of approximate measurement invariance tests compare to the results of the exact measurement invariance test is missing. We address this gap by comparing the results of exact and approximate (Bayesian) cross-country measurement invariance tests of the revised scale to measure human values. We query whether the approximate (more liberal) approach yields higher levels of measurement invariance for the values scale than the exact approach.

SCHWARTZ'S THEORY OF BASIC HUMAN VALUES

Schwartz (1992), Schwartz et al. (2012) defines values as broad, trans-situational goals that vary in importance and serve as guiding principles in the life of a person or group. Schwartz distinguishes between value hierarchies and value structure. Value hierarchies refer to the relative importance of the set of values to different individuals. The central claim of Schwartz's value theory concerns the value structure. It asserts that values form a circular motivational continuum. This means that values that are located in adjacent regions on the continuum are motivationally similar. Behavior that expresses one value is likely to express the adjacent values at the same time. In contrast, values that are located on opposing sides of the circle express conflicting motivations; hence, behavior that expresses one value is likely to simultaneously challenge or block the expression of opposing values in the circle.

The claim that values form a continuum implies that the circle of values can be partitioned in any number of ways. Depending on the aims of a study, one can differentiate between fewer broadly defined values or many more narrowly defined values. There are two common ways of partitioning the circular continuum, the classic version and the refined version. The classic version (Schwartz, 1992) partitions the circle into 10 basic human values. The refined version (Schwartz et al., 2012) partitions the circle into 19 more narrowly defined values. The 19 values in the refined version are subdimensions of the 10 basic human values (Schwartz et al., 2012). The values in both versions can be grouped into sets of four higher-order values: person-oriented vs. socially-oriented values or self-protection vs. growth values. Thus, the refined version of the theory and the classic version both describe the same circular motivational continuum. However, the refined theory provides a more discriminate partitioning of the continuum, thus allowing more fine-tuned predictions and explanations. **Figure 1** depicts the value circle with its 19 narrowly defined values, and the definition of each value is presented in **Table 1**.

MEASUREMENT OF BASIC HUMAN VALUES

The problem of measurement invariance is especially important for values because researchers often use them to describe differences between demographic, occupational, cultural, and national groups (Inglehart and Baker, 2000; Schwartz, 2006). Several methods have been developed to measure the values in Schwartz's

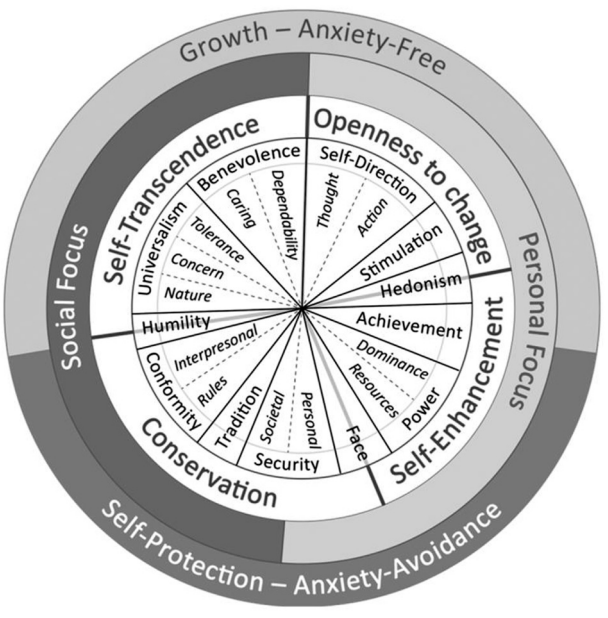


FIGURE 1 | The circular motivational continuum of 19 values in the refined value theory (Cieciuch et al., 2014).

Table 1 | Nineteen more narrowly defined values in the refined theory of values (Schwartz et al., 2012).

Value	Conceptual definitions in terms of motivational goals
Self-direction—Thought	Freedom to cultivate one's own ideas and abilities
Self-direction—Action	Freedom to determine one's own actions
Stimulation	Excitement, novelty, and change
Hedonism	Pleasure and sensuous gratification
Achievement	Success according to social standards
Power—Dominance	Power through exercising control over people
Power—Resources	Power through control of material and social resources
Face	Security and power through maintaining one's public image and avoiding humiliation
Security—Personal	Safety in one's immediate environment
Security—Societal	Safety and stability in the wider society
Tradition	Maintaining and preserving cultural, family, or religious traditions
Conformity—Rules	Compliance with rules, laws, and formal obligations
Conformity—Interpersonal	Avoidance of upsetting or harming other people
Humility	Recognizing one's insignificance in the larger scheme of things
Benevolence—Dependability	Being a reliable and trustworthy member of the ingroup
Benevolence—Caring	Devotion to the welfare of ingroup members
Universalism—Concern	Commitment to equality, justice, and protection for all people
Universalism—Nature	Preservation of the natural environment
Universalism—Tolerance	Acceptance and understanding of those who are different from oneself

approach. Currently, the most commonly used questionnaires are several versions of the Portrait Value Questionnaire (PVQ). The original version (PVQ-40) includes 40 items (Schwartz et al., 2001; Schwartz, 2003). A shorter version, implemented in the European Social Survey (ESS), includes 21 items (PVQ-21, Schwartz, 2003). The most recent version, developed to measure the 19 values of the refined value theory, includes 57 items (PVQ-57, Schwartz et al., 2012).

Several studies have tested the measurement invariance across countries of the PVQ-21 with data collected in the ESS (e.g., Davidov, 2008, 2010; Davidov et al., 2008). These studies succeeded in identifying only seven values at the configural level; it was necessary to unify some pairs of adjacent values in the confirmatory factor analyses. Davidov et al. (2008) established metric invariance for these seven values, but not scalar invariance. The lack of scalar invariance even for these seven was problematic because it meant that comparisons of means across cultures or countries may not be meaningful.

Cieciuch and Davidov (2012) addressed this problem when they compared the invariance properties between the PVQ-21 and PVQ-40 across Poland and Germany. They found that the PVQ-40 displayed a higher level of measurement invariance than the PVQ-21; it attained scalar invariance for all of the values except stimulation. They attributed the superiority of the PVQ-40 to the larger number of indicators available to measure the latent factors. With more items, the possibility of establishing partial scalar invariance increases. The reason for this is that, when establishing partial invariance, researchers need to identify at least two items with equal parameters across groups. When the number of indicators measuring a construct increases, chances also increase to identify two such items.

To measure all of the narrowly defined values that are differentiated in the refined theory, Schwartz et al. (2012) developed the PVQ-57. This version introduced three important changes compared to previous versions of the PVQ: (1) Single sentences were used for all items, replacing the two-sentence items of earlier versions. This avoided the dangers associated with double-barreled questions and improved overall clarity. (2) All items referred to the "importance" of a valued goal or characteristic to the respondent, replacing terms that referred to desires and feelings in earlier versions. This increase in consistency ensured that all items fit the conception of values as goals that vary in importance. (3) Three items measured each of the 19 values, which is in contrast to the varying number of items for each value in the PVQ-40 and the two items in the PVQ-21.

CFA analyses of the revised PVQ instrument successfully identified all 19 values in eight countries (Finland, Germany, Israel, Italy, New Zealand, Poland, Portugal, and Switzerland), establishing both configural and metric invariance (Cieciuch et al., 2014). Moreover, Cieciuch et al. (2014) established scalar measurement invariance for items measuring 10 of the 19 values across the eight countries. **Table 5** presents the detailed results of these analyses. Encouraging as these findings are in allowing comparison of means across countries for 10 values, a problem remains with the other nine values for which scalar invariance was not established. Perhaps, however, the method used to test measurement invariance test was overly strict. We therefore asked whether a more liberal test would yield more invariant results.

THE CURRENT STUDY

Several researchers have recently argued that, although measurement invariance is necessary for meaningful comparisons across groups, the criteria for evaluating measurement invariance are too strict (Muthén and Asparouhov, 2013; Van de Schoot et al., 2013; Muthén, 2014). This may lead to rejecting the possibility of comparison and needlessly discourage research in some cases. Adopting this view, Muthén and Asparouhov (2013) proposed the concept of approximate rather than exact measurement invariance, which is based on Bayesian analysis.

APPROXIMATE (BAYESIAN) MEASUREMENT INVARIANCE

Bayesian analysis allows researchers to introduce existing knowledge into their analyses, especially the amount of uncertainty. The current practice within the dominant frequentist approach is to use existing knowledge in the theoretical introduction of papers

and in the discussion but seldom in the analyses. Often the testing of null hypotheses ignores the existence of prior knowledge. Bayesian analysis allows testing informative hypotheses, that is, hypotheses that take prior knowledge into account. This logic may also be applied to testing measurement invariance.

In the Bayesian approach, parameters (e.g., loadings or intercepts) are considered to be variables with a specific distribution. The parameters of this distribution are called priors and can be defined by the researcher based on previous knowledge or assumptions (Muthén and Asparouhov, 2013). In the exact measurement invariance approach, researchers assume that the differences between loadings (or intercepts) across groups are zero or, in other words, that the loadings (or intercepts) are exactly equal across groups. The Bayesian measurement invariance approach introduces the concept of approximate equality. Thus, for testing approximate measurement invariance, one can expect that some differences in loadings (or intercepts) can occur, however, the mean of the differences between loadings (or intercepts) across groups is zero. Because the low variability is rather random, a normal distribution of the differences in loadings (or intercepts) with zero mean and small variance is assumed. Several simulation studies have shown that small variations (variance equal to 0.01 or 0.05) in the distribution of the differences in loadings or intercepts do not bias substantive conclusions for comparative research (Muthén and Asparouhov, 2013; Van de Schoot et al., 2013). Consequently, it makes sense to regard a small amount of variation as acceptable. Approximate measurement invariance differs from the partial measurement invariance approach, because in the latter some parameters are constrained to be exactly equal and others are released entirely, while in the former all parameters are constrained; however, the restrictions are more liberal and refer to the concept of approximate equality.

In the next section we test for approximate measurement invariance of the 19 values from the refined value theory of Schwartz et al. (2012). We then compare the findings to those established in previous studies that used exact measurement invariance testing.

Approximate measurement invariance is a relatively new approach. Therefore, there are few comparisons in the literature of the results that this approach yields with those obtained by the classic, exact measurement, invariance approach. We expect that the new scale to measure 19 values will exhibit a higher invariance level than the one reported by Cieciuch et al. (2014) when approximate measurement invariance is applied, because it allows for small differences between parameters that are otherwise constrained to be exactly equal in the exact measurement invariance approach. This would justify doing additional cross-cultural comparisons.

METHODS

PARTICIPANTS AND PROCEDURE

We used the same data employed for testing exact measurement invariance in Cieciuch et al. (2014). Data were from the following countries: Finland ($N = 334$, 65% female, $M_{age} = 42.3$, $SD_{age} = 6.1$), Germany ($N = 325$, 77% female, $M_{age} = 23.4$, $SD_{age} = 5.0$), Israel ($N = 394$, 65% female, $M_{age} = 25.7$, $SD_{age} = 6.2$), Italy ($N = 388$, 59% female, $M_{age} = 35.6$, $SD_{age} = 14.5$), New Zealand ($N = 527$, 68% female, $M_{age} = 19.5$, $SD_{age} = 4.2$), Poland ($N = 547$, 66% female, $M_{age} = 27.0$, $SD_{age} = 10.0$), Portugal ($N = 295$, 58% female, $M_{age} = 27.0$, $SD_{age} = 10.4$), and Switzerland ($N = 201$, 70% female, $M_{age} = 28.8$, $SD_{age} = 7.7$). All participants were contacted by researchers or instructed assistants in person or online and completed the value instrument voluntarily and anonymously. Data were collected in a written format in Finland, Germany, Italy, Poland, and in half the Portuguese sample. Data were collected online in the remaining samples. All data are available from the first author upon request.

QUESTIONNAIRE

Data were collected with the PVQ-5X (Schwartz et al., 2012) developed to measure 19 more narrowly defined values. Items described a person in terms of what is important for him or her (gender matched). The respondents were asked to answer the question "*How much is this person like you*" on a scale ranging from 1 (*not like me at all*) to 6 (*very much like me*). For example, the question "Freedom to choose what he does is important to him" measured the self-direction value. The question "Obeying all the laws is important to her" was used to measure the value conformity rules. All items are presented in **Table 4**. We excluded nine items which did not load satisfactorily on their corresponding value in the study of Schwartz et al. (2012). Thus, our analyses included exactly the same items included in the exact measurement invariance test of Cieciuch et al. (2014). Ten of the values were measured by three indicators and nine values by two indicators. Missing values for all items were below 0.7% with the exception of one achievement item (AC1) which had 2.9% missing values.

ANALYSIS

TESTING FOR APPROXIMATE MEASUREMENT INVARIANCE IN Mplus (VERSION 7.11)

The approximate measurement invariance test procedure is included in Mplus (Muthén and Muthén, 1998–2012) in the mixture analysis framework. Mixture modeling means that besides the latent variables included in the model, there are also one or more latent categorical variables that describe membership of respondents to a certain class. These latent categorical variables represent homogenous subpopulations of the studied heterogeneous population (Muthén, 2002). In principle, mixture modeling assumes that the division into subpopulations and subpopulation membership are not known but can be inferred from the data. However, in our case this was a straightforward inference, because the population membership was deduced by the country where data on the individuals were collected. Thus, this categorical variable was known, since it was simply the variable that described membership in groups (countries). In terms of mixture models, this situation is known as a single-class mixture model because there is only one class (one categorical variable). According to Asparouhov and Muthén (2010), if the categorical variable is observed, the single-class mixture model is essentially the same as a multigroup model. Kim et al. (2013) also argue that the two models (i.e., the multigroup model and the single-class mixture model with known class membership) are in principle the same.

Table 2 presents the syntax, briefly explains the various steps of the analysis, and provides a description of the statements used in the syntax.

EVALUATION OF THE MODEL

The fit of the Bayesian model can detect whether actual deviations are larger than those that the researcher allows in the prior distribution. The model fit can be evaluated based on the posterior predictive probability (ppp) value and the confidence interval (CI) for the difference between the observed and replicated chi-square values. According to Muthén and Asparouhov (2013) and Van de Schoot et al. (2013), the Bayesian model fits the data when the ppp is higher than zero[1] and the CI contains zero. We defined the mean of the differences in loadings and intercepts across countries as zero and the variance of these differences as 0.01 (Van de Schoot et al., 2013). If the model was unacceptable based on the ppp and the CI, we slightly increased the variance to determine the level of variation in the priors for the difference between loadings and intercepts that would lead to acceptable model fit coefficients[2]. Additionally, Mplus lists all parameters that significantly differ from the priors. This feature is equivalent to modification indices in the exact measurement invariance approach. While the model is assessed based on ppp and CI, these values provide global model fit criteria that are similar to the criteria in the exact approach (Chen, 2007). Although several parameters have been identified as exactly equal in Cieciuch et al. (2014), we did not constrain them to equality and allowed a wiggle for the differences between all factor loadings and intercepts. The reason is that we wanted to assess whether a liberal model would establish invariance for all values.

RESULTS

Table 3 presents the fit coefficients of the approximate multigroup CFA for each value separately. For most of the values, the ppp was not significant, and the 95% CI for the difference between the observed and replicated chi-square values contained zero, which means that the approximate scalar invariance models for these values are acceptable. The only three exceptions were stimulation, achievement, and humility. Therefore, we increased the variance prior for these values to 0.02. With this adjustment, all three approximate scalar invariance models were also acceptable for these values. In other words, the model fit criteria suggest that approximate invariance could be established for all 19 values across eight countries.

Several loadings and intercepts in various countries deviated from the defined priors. For example, the intercept of the first item measuring Self-direction–Thought (SDT1) deviated from the defined prior in two countries, Finland and Poland. The loading of the first item measuring Stimulation (ST1) deviated in two countries, Italy and Poland, and its intercept deviated from the defined prior in two countries as well, Italy and New Zealand. **Table 4** presents all deviations of loadings and intercepts from the

[1]Simulation studies are still required to determine what level of probability researchers may rely on.

[2]There are still no established cut-off criteria in the literature about the maximal level of variability that may be used for the priors.

Table 2 | Mplus syntax for approximate measurement invariance test and explanations (this is an example for a single factor—UNC).

VARIABLE:	
Names are country UNC1 UNC2 UNC3;	This indicates the variables in the data: the countries and the items for each value (Universalism-concern in this example).
classes = c(8);	This option specifies that there is one latent categorical variable (named c) that has 8 latent classes. The number 8 refers to 8 countries in the analysis.
knownclass = c(country = 1 2 3 4 5 6 7 8);	This option defines the categorical latent variable by the observed variable. There are 8 classes and respondents with value 1 in variable "country" belong to the first one; respondents with value 2 in variable "country" belongs to the second country, etc. If all values from the variable are to be analyzed, the statement can be shortened: knownclass = c (country).
ANALYSIS:	
type = mixture;	Approximate measurement invariance is included in Mplus within the mixture modeling analysis framework. The number of classes is known because it corresponds to the number of groups to be compared.
Estimator = bayes;	Bayesian analysis will be performed and priors can be defined.
chains is 5;	The number of chains in Markov chain Monte Carlo (MCMC) algorithms. The default in Mplus is 2 chains and the researcher can increase the number of chains by this statement.
Processor = 5;	To increase the speed of computation, one can use more processors if they are available in the hardware. It is possible to specify the number of processors that is equal to number of chains. In this case one can specify also 8 processors. If that many processors are not available, each available processor carries out one chain and after it is completed starts with the next chain.
Biterations = 500,000(20,000);	This option is used to specify the maximum and minimum number of iterations for each Markov chain Monte Carlo algorithm. In this case, it specifies that a minimum of 20,000 and a maximum of 50,000 iterations will be used.
Bconvergence = 0.01;	Specification of the convergence value criterion to be used for determining convergence of the Bayesian estimation.
bseed 100;	Specification of the seed to be used for a random number generation in the Markov chain Monte Carlo (MCMC) algorithm (the default in Mplus is zero).
model = allfree;	Factor means, variances, and covariances are freely estimated across groups with the exception of factor means in the last group which are fixed to 0.

(Continued)

Table 2 | Continued

MODEL:	
%overall% UNC by UNC1* UNC 2 UNC 3 (lam#_1-lam#_3); [UNC 1 UNC 2 UNC 3] (nu#_1-nu#_3);	In the mixture models, the label "%overall%" introduces the model description which is common for all groups. In this case the latent variable is loaded by three indicators (UNC1, UNC2, and UNC3). The asterisk after UNC1 implies that the loadings of the first indicator, which is usually constrained by default to 1, is freed. Following the "by" statement, the names of the factor loadings are listed in parentheses. One row below, after the brackets, the names of the intercepts are listed. It is necessary to list these so that one can later define their priors.
MODEL PRIORS:	
do(1,3) diff(lam1_#-lam8_#)∼N(0,0.01); do(1,3) diff(nu1_#-nu8_#)∼N(0,0.01);	The statement defines priors for loadings and intercepts. The distribution of loadings and intercepts is normal with mean = 0 and variance = 0.01
%c#8% [UNC @0]; UNC @1;	The label "%c#8%" refers to the part of the model for class 8 that differs from the overall model. In this case, the latent mean of UNC in the last group is constrained to 0 and the variance to 1 in order to identify the model according to the proposal of Muthén and Asparouhov (2013).

Table 3 | Model fit coefficients of Bayesian multigroup confirmatory factor analysis for each value.

	ppp	95% *CI*
Self-direction—Thought	0.201	(−19.478) – (49.818)
Self-direction—Action	0.112	(−12.931) – (57.474)
Stimulation	0.001	(25.824) – (110.628)
Stimulation, prior of variance = 0.02	0.081	(−9.495) – (64.259)
Hedonism	0.258	(−18.255) – (35.833)
Achievement	0.004	(20.132) – (98.707)
Achievement, prior of variance = 0.02	0.103	(−13.481) – (62.092)
Power—Resources	0.367	(−22.056) – (30.480)
Power—Dominance	0.208	(−15.653) – (37.917)
Face*	0.128	(−11.916) – (45.275)
Security—Personal	0.361	(−20.384) – (32.179)
Security—Societal	0.135	(−13.923) – (55.015)
Tradition	0.028	(−0.594) – (76.570)
Conformity—Rules	0.352	(−20.444) – (30.633)
Conformity—Interpersonal	0.083	(−11.226) – (65.544)
Humility*	0.009	(6.575) – (70.861)
Humility, prior of variance = 0.02	0.121	(−11.877) – (46.340)
Benevolence—Caring	0.506	(−34.843) – (33.737)
Benevolence—Dependability*	0.149	(−12.476) – (43.798)
Universalism—Concern	0.235	(−25.179) – (47.297)
Universalism—Nature	0.167	(−18.021) – (51.002)
Universalism—Tolerance	0.395	(−23.183) – (31.304)

*ppp = posterior predictive p-value; 95% CI = Confidence interval for the difference between the observed and the replicated chi-square values, *because of estimation problems, the latent means were constrained to 0 and variances to 1 in two countries for this value rather than in one country. These additional constraints were not rejected by the model.*

defined priors. Despite the deviations listed in **Table 4**, the ppp and CI reached acceptable levels, which suggests that approximate metric and scalar measurement invariance are supported by the data for all values.

Table 5 presents a comparison of Cieciuch et al.'s (2014) results using the exact approach and the results in the current study obtained using the approximate approach. Whereas exact scalar invariance was previously supported only for a subset of the 19 values, in the present analysis, approximate measurement invariance was established for all values, including those values where exact measurement invariance testing failed to display scalar invariance. In the next section we are going to discuss in more detail the results, their implications, and limitations.

SUMMARY AND CONCLUSIONS

Measurement invariance is a precondition for meaningful cross-group comparisons. Assuming rather than empirically testing whether the precondition is satisfied can be dangerous and can lead to wrong conclusions. Therefore, an empirical test of measurement invariance of a study's measures is necessary. However, the classic (exact) test is very demanding and very often leads to the rejection of measurement invariance and to precluding group comparisons. Van de Schoot et al. (2013) metaphorically described this situation as traveling between Scylla and Charybdis. Scylla represents the situation in which a model lacks measurement invariance, whereas Charybdis represents the situation in which the model was not tested for measurement invariance. In both situations, the researcher cannot know whether the differences between groups are real and substantive or a result of methodological artifacts. We followed Van de Schoot et al. (2013) suggestion to choose a third option for traveling between Scylla and Charybdis. This option is the approximate Bayesian approach to measurement invariance. Approximate measurement invariance is a rather new approach and applications using it and comparing its findings to those of the exact approach are rare. Using data on human values in eight countries, we tried to fill this gap by comparing the findings of an earlier analysis using the exact approach to measurement invariance by analyzing the same data using the approximate approach.

The approximate approach established measurement invariance across eight countries for the new PVQ-5X scale to measure human values even in cases in which the exact approach did not. In other words, the approximate method is less restrictive than the exact, and our findings suggest that—as expected—the results align with this, i.e., the less restrictive method (approximate invariance testing using the Bayesian procedure) produces stronger invariance than the exact approach did. These findings provide, for the first time, initial encouraging results that the PVQ-5X scale may be used for conducting meaningful cross-cultural research with all 19 values. The exact approach to assessing invariance has often shed doubt on the invariance of many questionnaires. The current findings provide hope that empirical

Table 4 | Deviations of loadings and intercepts from prior defined parameters (mean = 0, variance = 0.01).

	Finland		Israel		Italy		New Zealand		Poland		Portugal		Switzerland		Germany	
	Lo	Int	Lo	Int	Lo	Int	Lo	Int	Lo	Int	Lo	Int	Lo	Int	Lo	Int
SDT1 Being creative is important to him		x								x						
SDT2 It is important to him to form his own opinions and have original ideas						x										
SDT3 Learning things for himself and improving his abilities is important to him		x				x				x		x				
SDA1 It is important to him to make his own decisions about his life	x	x				x		x								
SDA2 Doing everything independently is important to him		x				x		x		x					x	
SDA3 Freedom to choose what he does is important to him					x	x										
ST1 He is always looking for different kinds of things to do					x	x		x	x							
ST2 Excitement in life is important to him			x						x			x		x		x
ST3 He thinks it is important to have all sorts of new experiences		x				x						x	x	x		
HE1 Having a good time is important to him						x			x	x						
HE2 Enjoying life's pleasures is important to him																
AC1 He thinks it is important to be ambitious				x					x	x		x				x
AC2 Being very successful is important to him																
AC3 He wants people to admire his achievements				x						x		x				x
POR1 Having the feeling of power that money can bring is important to him																
POR2 Being wealthy is important to him																
POD1 He wants people to do what he says						x										
POD3 It is important to him to be the one who tells others what to do						x										
FAC1 It is important to him that no one should ever shame him										x						
FAC2 Protecting his public image is important to him										x						
SEP2 His personal security is extremely important to him																
SEP3 It is important to him to live in secure surroundings																
SES1 It is important to him that his country protect itself against all threats																
SES2 He wants the state to be strong so it can defend its citizens										x						
SES3 Having order and stability in society is important to him		x	x	x					x	x						

(Continued)

Table 4 | Continued

	Finland		Israel		Italy		New Zealand		Poland		Portugal		Switzerland		Germany	
	Lo	**Int**	**Lo**	**Int**	**Lo**	**Int**	**Lo**	**Int**	**Lo**	**Int**	**Lo**	**Int**	**Lo**	**Int**	**Lo**	**Int**
TR1 It is important to him to maintain traditional values or beliefs		x		x		x		x		x				x		x
TR2 Following his family's customs or the customs of a religion is important to him		x								x				x		x
TR3 He strongly values the traditional practices of his culture												x				
COR2 It is important to him to follow rules even when no one is watching								x								
COR3 Obeying all the laws is important to him																
COI1 It is important to him to avoid upsetting other people		x			x	x		x	x	x		x				
COI2 He thinks it is important never to be annoying to anyone				x		x		x		x						
COI3 He always tries to be tactful and avoid irritating people				x		x										x
HU2 It is important to him to be humble																
HU3 It is important to him to be satisfied with what he has and not to ask for more																
BEC1 It's very important to him to help the people dear to him															x	
BEC2 Caring for the well-being of people he is close to is important to him												x			x	
BEC3 (BED1) it is important to him to be loyal to those who are close to him					x											
BED2 He goes out of his way to be a dependable and trustworthy friend							x	x								
BED3 He wants those he spends time with to be able to rely on him completely	x			x					x	x			x	x		x
UNC1 Protecting society's weak and vulnerable members is important to him								x								x
UNC2 He thinks it is important that every person in the world have equal opportunities in life																
UNC3 He wants everyone to be treated justly, even people he doesn't know								x				x			x	x
UNN1 He strongly believes that he should care for nature		x		x		x		x							x	
UNN2 It is important to him to work against threats to the world of nature		x										x				
UNN3 Protecting the natural environment from destruction or pollution is important to him								x								
UNT2 It is important to him to listen to people who are different from him																
UNT3 Even when he disagrees with people, it is important to him to understand them																

Lo = loading; Int = intercept; x—deviation of a given parameter in a given group from the defined priors (mean = 0, variance = 0.01).

Table 5 | Comparison of exact and approximate measurement invariance of 19 values across eight countries.

	Exact (Cieciuch et al., 2014)		Approximate (the current study)
	Metric	Scalar	Metric and scalar
Self-direction thought	Full in all countries	Partial in all countries	Full in all countries
Self-direction action	Full in five countries, partial in Finland and Portugal, absent in Italy	Full in all countries	Full in all countries
Stimulation	Full in all countries	Full in all countries	Full in all countries*
Hedonism	Full in seven countries, Absent in Switzerland	Full in six countries, absent in Switzerland, Poland	Full in all countries
Achievement	Full in six countries, partial in Finland and Poland	Absent in all countries	Full in all countries*
Power dominance	Full in all countries	Full in six countries, absent in Portugal, Italy	Full in all countries
Power resources	Full in all countries	Full in seven countries, absent in Poland	Full in all countries
Face	Full in all countries	Absent in all countries	Full in all countries
Security personal	Full in all countries	Full in six countries, absent in Israel and Switzerland	Full in all countries
Societal security	Full in seven countries, partial in Portugal	Partial in all countries	Full in all countries
Tradition	Full in all countries	Absent in all countries	Full in all countries
Conformity rules	Full in all countries	Absent in all countries	Full in all countries
Conformity interpersonal	Full in all countries	Absent in all countries	Full in all countries
Humility	Full in all countries	Absent in all countries	Full in all countries*
Universalism nature	Full in all countries	Full in four countries, partial in Israel, Italy, and New Zealand, absent in Switzerland	Full in all countries
Universalism concern	Full in all countries	Full in five countries, partial in New Zealand, Portugal, absent in Germany	Full in all countries
Universalism tolerance	Full in all countries	Full in six countries, absent in Poland and Portugal	Full in all countries
Benevolence caring	Full in all countries	Full in seven countries, partial in Finland	Full in all countries
Benevolence dependability	Full in all countries	Absent in all countries	Full in all countries

**The allowed variance for the cross-country difference between intercepts and the loadings was 0.02. In all other cases it was 0.01.*

testing for measurement invariance in questionnaires is not necessarily doomed to failure. Researchers may now put their scales to even a stricter test and examine whether some of the parameters may be constrained to be exactly (rather than approximately) equal.

Findings raise the question whether other established scales to measure human values such as the PVQ-21 scale included in the ESS will display higher levels of equivalence across countries when using the approximate Bayesian (rather than an exact) approach for the test. Future research should address this question by investigating the cross-country comparability of other scales to measure human values using the Bayesian approximate invariance approach.

This study is not without limitations. First, we used convenience student samples and data were collected using different modes of data collection (online and offline). Although previous studies (e.g., Davidov and Depner, 2011) demonstrated that online and offline modes of data collection produce invariant value measurements, future studies should address this issue by trying to validate and generalize our findings using country population samples. Second, we do not know whether and to what extent the different sample sizes across countries (e.g., 547 in Poland vs. 201 in Switzerland) may have disproportionally biased the fit measures. In his simulations, Chen (2007) provided recommendations for model fit evaluation for different sample sizes when testing for exact measurement invariance. However, we are not aware of any such simulations for the Bayesian approach. Future research should address the robustness of the model fit criteria to different sample sizes. Furthermore, it is not clear whether and to what extent the fact that the outcomes are ordinal might affect the results. Whereas exact measurement invariance tests can take the ordinal character of item scores into account in the estimation, unfortunately, the Bayesian approach does not deal with this problem appropriately and assumes that scores are continuous. We can only speculate that this may bias our conclusions but it is difficult to judge in which direction. Future research should address this problem by developing Bayesian procedures that allow testing for approximate measurement invariance while taking into account the ordinal character of the data. Yet it should be noted that our response scale included six categories, one more than the common five-point Likert scales, so this should have hopefully mitigated the problem.

In spite of our encouraging findings, an important unanswered question remains to be resolved: What is the magnitude of the variance that should be specified for the priors? Specifying a small variance may result in failure to establish invariance while specifying a larger variance may lead to establishing invariance. We set a magnitude of 0.01 and in three cases increased it to 0.02 in order to establish invariance. These seem like small magnitudes, but are they too liberal? This technical question is extremely important from an applied point of view. Finally, it is too early to claim that researchers should now switch to testing for approximate measurement invariance (instead of testing for exact measurement invariance). It is still a rather unexplored field, and further studies are needed before such a claim can be fully justified. In addition to the promising results reported here, further research and simulation studies should focus on these questions to provide guidelines for applied researchers.

ACKNOWLEDGMENTS

The work of the first, the second, and the fourth authors was supported by the University Research Priority Program 'Social Networks' of the University of Zürich. The work of the first author was partially supported by Grant 2011/01/D/HS6/04077 from the Polish National Science Center. The authors would like to thank Lisa Trierweiler for the English proof of the manuscript.

REFERENCES

Asparouhov, T., and Muthén, B. O. (2010). *Bayesian Analysis Using Mplus: Technical Implementation*. Available online at: http://www.statmodel.com/download/Bayes3.pdf

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: Wiley

Byrne, B. M., Shavelson, R. J., and Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures–the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling* 14, 464–504. doi: 10.1080/10705510701301834

Cieciuch, J., and Davidov, E. (2012). A comparison of the invariance properties of the PVQ-40 and the PVQ-21 to measure human values across German and Polish samples. *Surv. Res. Method* 6, 37–48.

Cieciuch, J., Davidov, E., Vecchione, M., Beierlein, C., and Schwartz, S. H. (2014). The cross-national invariance properties of a new scale to measure 19 basic human values. A test across eight countries. *J. Cross Cult. Psychol.* 45, 764–779. doi: 10.1177/0022022114527348

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

Davidov, E. (2010). Testing for comparability of human values across countries and time with the third round of the European Social Survey. *Int. J. Comp. Sociol.* 51, 171–191. doi: 10.1177/0020715210363534

Davidov, E., and Depner, F. (2011). Testing for measurement equivalence of human values across online and paper-and-pencil surveys. *Qual. Quant.* 45, 375–390. doi: 10.1007/s11135-009-9297-9

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., and Billiet, J. (2014). Measurement equivalence in cross-national research. *Annu. Rev. Sociol.* 40, 55–75. doi: 10.1146/annurev-soc-071913-043137

Davidov, E., Schmidt, P., and Schwartz, S. (2008). Bringing values back in. The adequacy of the European Social Survey to measure values in 20 countries. *Public Opin. Q.* 72, 420–445. doi: 10.1093/poq/nfn035

Horn, J. L., and McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Exp. Aging Res.* 18, 117–144. doi: 10.1080/03610739208253916

Inglehart, R., and Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *Am. Sociol. Rev.* 65, 19–51. doi: 10.2307/2657288

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika* 36, 409–426. doi: 10.1007/BF02291366

Kim, S. Y., Mun, E. Y., and Smith, S. (2013). Using mixture models with known class membership to address incomplete covariance structures in multiple-group growth models. *Br. J. Math. Stat. Psychol.* 67, 94–116. doi: 10.1111/bmsp.12008

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Muthén, B. O. (2002). Beyond SEM: general latent variable modeling. *Behaviormetrika* 29, 81–117. doi: 10.2333/bhmk.29.81

Muthén, B. O. (2014). IRT studies of many groups: the alignment method. *Front. Psychol.* 5:978. doi: 10.3389/fpsyg.2014.00978

Muthén, B. O., and Asparouhov, T. (2013). *BSEM Measurement Invariance Analysis. Mplus Web Notes: No. 17*. Available online at: www.statmodel.com

Muthén, L., and Muthén, B. O. (1998–2012). *Mplus User's Guide, 7th Edn*. Los Angeles, CA: Muthén and Muthén.

Schwartz, S. H. (1992). "Universals in the content and structure of values: theory and empirical tests in 20 countries," in *Advances in Experimental Social Psychology, Vol. 25*, ed M. Zanna (New York, NY: Academic Press), 1–65.

Schwartz, S. H. (2003). "A proposal for measuring value orientations across nations," in *Questionnaire Development Package of the European Social Survey, Chapter 7*. Available online at: www.europeansocialsurvey.org

Schwartz, S. H. (2006). A theory of cultural value orientations: explication and applications. *Comp. Sociol.* 5, 137–182. doi: 10.1163/156913306778667357

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., et al. (2012). Refining the theory of basic individual values. *J. Pers. Soc. Psychol.* 103, 663–688. doi: 10.1037/a0029393

Schwartz, S. H., Melech, G., Lehmann, A., Burgess, S., and Harris, M. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *J. Cross Cult. Psychol.* 32, 519–542. doi: 10.1177/0022022101032005001

Steenkamp, J.-B. E. M., and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *J. Consum. Res.* 25, 78–90. doi: 10.1086/209528

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthén, B. O. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organ. Res. Methods* 5, 139–158. doi: 10.1177/1094428102005002001

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 May 2014; accepted: 19 August 2014; published online: 08 September 2014.

*Citation: Cieciuch J, Davidov E, Schmidt P, Algesheimer R and Schwartz SH (2014) Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. Front. Psychol. **5**:982. doi: 10.3389/fpsyg.2014.00982*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

METHODS ARTICLE
published: 20 June 2014
doi: 10.3389/fpsyg.2014.00604

What's hampering measurement invariance: detecting non-invariant items using clusterwise simultaneous component analysis

Kim De Roover[1]*, Marieke E. Timmerman[2], Jozefien De Leersnyder[1], Batja Mesquita[1] and Eva Ceulemans[1]

[1] *Methods, Individual and Cultural Differences, Affect and Social Behavior, KU Leuven, Leuven, Belgium*
[2] *Heymans Institute of Psychology, University of Groningen, Groningen, Netherlands*

Edited by:
Rens Van De Schoot, Utrecht University, Netherlands

Reviewed by:
Heungsun Hwang, McGill University, Canada
P. M. Kroonenberg, Leiden University, Netherlands

***Correspondence:**
Kim De Roover, Methodology of Educational Sciences Research Group, Andreas Vesaliusstraat 2, Leuven B-3000, Belgium
e-mail: kim.deroover@ppw.kuleuven.be

The issue of measurement invariance is ubiquitous in the behavioral sciences nowadays as more and more studies yield multivariate multigroup data. When measurement invariance cannot be established across groups, this is often due to different loadings on only a few items. Within the multigroup CFA framework, methods have been proposed to trace such non-invariant items, but these methods have some disadvantages in that they require researchers to run a multitude of analyses and in that they imply assumptions that are often questionable. In this paper, we propose an alternative strategy which builds on clusterwise simultaneous component analysis (SCA). Clusterwise SCA, being an exploratory technique, assigns the groups under study to a few clusters based on differences and similarities in the component structure of the items, and thus based on the covariance matrices. Non-invariant items can then be traced by comparing the cluster-specific component loadings via congruence coefficients, which is far more parsimonious than comparing the component structure of all separate groups. In this paper we present a heuristic for this procedure. Afterwards, one can return to the multigroup CFA framework and check whether removing the non-invariant items or removing some of the equality restrictions for these items, yields satisfactory invariance test results. An empirical application concerning cross-cultural emotion data is used to demonstrate that this novel approach is useful and can co-exist with the traditional CFA approaches.

Keywords: measurement bias, configural invariance, weak invariance, metric invariance

INTRODUCTION

To assess the quality of psychological instruments (e.g., surveys, questionnaires, etc.), confirmatory factor analysis (CFA; Lawley and Maxwell, 1962) is often applied. CFA tests whether or not a particular latent variable model, specifying which latent variables (i.e., factors) are measured by which items, complies with the observed item scores. When the instrument is used among several groups, quality testing becomes more intricate, as the equality of different aspects of the latent variable model has to be verified (i.e., the configuration and size of the loadings of the items on the factors, item intercepts, unique variances), before the factor scores of the different groups can be compared meaningfully. For instance, when investigating cross-cultural differences in emotional experience, one has to make sure that the items of the emotion questionnaire behave the same across cultural groups. The different tests involved pertain to different levels of measurement invariance (Meredith, 1993; Meredith and Teresi, 2006) and can be performed using multigroup CFA (Jöreskog, 1971; Sörbom, 1974). In this paper, we propose a new procedure to detect which items violate configural and/or weak measurement invariance. Thus, we focus on equality of within-group covariance structures and do not consider invariance of intercepts or unique variances, or structural invariance (i.e., invariance of factor means, variances, and covariances). The novel procedure is rooted in component analysis[1] and circumvents some disadvantages of the existing solutions in the multigroup CFA framework.

Configural invariance, which usually is the baseline model in invariance testing, implies that the same number of factors and the same pattern of zero and free loadings is imposed in all groups. The configural invariance test examines whether the items are associated with the same factors in all groups or, in other words, whether the same latent variables are measured across the groups. Weak invariance (also referred to as "metric invariance") additionally investigates between-group agreement in how these latent variables are manifested. Specifically, it tests whether all factor loadings are equal across groups.

Traditionally, measurement invariance testing relied on conducting likelihood ratio tests (LRT) to evaluate whether adding

[1] Although the theoretical comparability of component and factor analyses has been heavily debated (e.g., Gorsuch, 1990), Velicer and Jackson (1990) established that component and factor loadings usually are highly similar and lead to the same interpretations with respect to underlying constructs.

invariance constraints caused a significant difference in the χ^2 fit statistics. This approach has two drawbacks, however. First, its performance heavily depends on sample size (Brannick, 1995; Kelloway, 1995). Second, in large samples even tiny violations, that are not interesting from a substantive point of view, result in a rejection of measurement invariance (Note that this is exactly what a hypothesis test ought to do). To circumvent the two drawbacks associated with LRT testing, alternative goodness-of-fit indices, such as the comparative fit index (CFI; Bentler, 1990) and the root mean square error of approximation (RMSEA; Steiger, 1989), have been developed. Criteria have been proposed for deciding whether these fit indices indicate good fit (Bentler, 1990; Hu and Bentler, 1999; Tabachnick and Fidell, 2005) and whether changes in these fit indices are meaningful or "practically significant" in the context of measurement invariance (Cheung and Rensvold, 2002). Throughout this paper, following Cheung and Rensvold (2002), we will use the CFI and consider a multigroup CFA model to have a good fit when the CFI is larger than 0.95 and a more constrained model to have a "significantly" worse fit than a less constrained model when the difference in CFI (ΔCFI) is larger than 0.01.

When configural and/or weak invariance cannot be established, different latent variables appear to be measured across the groups (i.e., no configural invariance) or the same latent variables are measured differently in these groups (i.e., no weak invariance), implying that factor scores cannot be sensibly compared across groups (note that the computation of factor scores has been vastly debated; e.g., Green, 1976; Gorsuch, 1983; Grice, 2001). In the multigroup CFA framework some solutions to this problem have been proposed, which aim at detecting which restrictions on the factor loadings should be removed.

A popular strategy[2] is the sequential model modification procedure (MacCallum, 1986; MacCallum et al., 1992), which uses modification indices to assess whether in specific groups secondary loadings are needed for some items (to solve the lack of configural invariance) and/or to detect which loadings should be allowed to vary across groups in the weak invariance model (leading to partial weak invariance; Byrne et al., 1989); such modifications are implemented one by one. A disadvantage of this method is that in each step of the procedure, the calculation of the modification indices is based on the assumption that all other loadings (except for the ones that were deemed to be non-invariant in the previous modification steps) are invariant. When this is not the case, the modification indices are inaccurate and may lead to incorrect modifications (Williams and Thomson, 1986; Cheung and Rensvold, 1999). Also, progressively modifying the factor model until it fits the data of all groups, increases the risk of capitalization on chance (MacCallum et al., 1992; Stuive et al., 2009).

Another strategy for dealing with violations of weak measurement invariance, is item-level invariance testing (Cheung and Rensvold, 1999). Assuming configural invariance, this method first checks whether some of the factors are non-invariant with respect to their loadings. Next, it examines for each of the n non-zero loadings on a non-invariant factor whether or not it can be restricted to be equal across groups. This entails conducting $n(n-1)/2$ invariance tests (i.e., one for each non-redundant combination of an invariant item and a reference item[3]) per non-invariant factor and integrating the results of these tests by means of a "triangle" heuristic. Specifically, an item is considered to be invariant with respect to the factor in question if restricting its loading to be equal across groups yields a CFI decrease smaller than 0.01, whichever of the other invariant items is used as a reference item (for more details, see Cheung and Rensvold, 1999).

Finally, Byrne and van de Vijver (2010) propose to delete all items one by one and to re-evaluate each time the goodness-of-fit of the multigroup CFA model. An item is flagged as non-invariant when its deletion causes the CFI to increase more than 0.01.

All three strategies become cumbersome if the number of items grows larger, because they are prone to chance-capitalization and are computationally demanding, and because their validity stands or falls with the validity of some stringent assumptions. Hence, although CFA solutions exist and are often used, these solutions are not without problems.

In this paper, we propose an alternative procedure for detecting items that are non-invariant with respect to the structure or size of their factor loadings. Our procedure circumvents some disadvantages of the CFA solutions in that it is fast and does not entail assumptions with respect to the invariance of certain items or loadings. It builds on the results of a clusterwise simultaneous component analysis (SCA; De Roover et al., 2012). Being an exploratory technique, clusterwise SCA assigns the groups under study to a few clusters based on differences and similarities in the component structure and thus in the covariance matrices of the items. Next, non-invariant items can be traced by comparing the cluster-specific component loadings (which is far more parsimonious than comparing the component structure of all separate groups). To do this in a consistent way, we present a heuristic that is based on the Tucker's congruence coefficient (Tucker, 1951), an index that is often used in, amongst others, cross-cultural psychology, to make statements about the similarity of group-specific factor structures (Lorenzo-Seva and ten Berge, 2006). Afterwards, one can return to the multigroup CFA framework and check whether removing the non-invariant items or removing some of the equality restrictions for these items, yields satisfactory invariance test results.

Clustering the groups based on their component structure is a unique feature of our approach, that makes it especially appealing when the number of groups is large. Indeed, in such cases the clustering parsimoniously reveals the most important structural differences whereas the CFA solutions discussed above quickly become very tedious and impractical. Vice versa, when the data comprise only a few groups, it makes less sense to cluster the groups and the traditional approaches may be preferred.

The remainder of this paper is organized into three sections: in the Methods section, we introduce some notation regarding the

[2]In this paper we focus on the frequentist framework when discussing different methods to investigate the lack of invariance. Note that also in the Bayesian framework, methods exist with a similar aim (e.g., Muthén and Asparouhov, 2012, 2013).

[3]The item for which the factor loading is fixed to one in each population for model identification.

data and discuss preprocessing. Next, we recapitulate clusterwise SCA and present the heuristic for the detection of non-invariant items. Then, the Applications section illustrates the procedure using an empirical data set from research on emotional acculturation including emotional patterns from 13 different cultural groups. Finally, the Discussion will address some limitations and strengths of the presented method as well as directions for future research.

METHODS

DATA

In this paper we will be working with multivariate multigroup data, consisting of a N_k (subjects) $\times$ J (items) data matrix $\mathbf{X}_k$ ($k = 1, \ldots, K$) for each of the K groups under study. Since clusterwise SCA aims to cluster the groups based on the within-group component structure and not on differences in group-specific item means, it is essential that the data of each group are centered per item. Moreover, since items with a higher amount of variance may dominate the obtained components, it will often be wise to rescale the data to eliminate differences between the items in measurement scale or variability[4]. As configural and weak invariance pertain to the covariance structures of the groups, we advocate to normalize the items over all groups, implying that (co)variance differences among the groups are retained in the data. That is, we recommend to analyze the $\mathbf{X}_k$ matrices, computed from the raw (i.e., unpreprocessed) data matrices $\mathbf{X}_k^{\mathrm{r}}$ as follows:

$$\mathbf{X}_k = (\mathbf{X}_k^{\mathrm{r}} - \mathbf{1}_{N_k}\bar{\mathbf{x}}_k)\mathbf{S}^{-1} \tag{1}$$

where $\mathbf{1}_k$ is a $K \times 1$ vector of ones, $\bar{\mathbf{x}}_k$ is a $1 \times J$ vector containing the group-specific item means, $\mathbf{S}$ is a diagonal matrix containing the standard deviations of the items over all groups.

CLUSTERWISE SCA-P

Simultaneous component analysis (SCA; Kiers and ten Berge, 1994; Timmerman and Kiers, 2003) reduces the data of all groups simultaneously, summarizing the observed items by means of a few components according to the item covariances. SCA assumes that the same components underlie the data of the different groups and thus that the same loading matrix can be used for all groups. Specifically, the SCA model is given by:

$$\mathbf{X}_k = \mathbf{F}_k\mathbf{B}' + \mathbf{E}_k \tag{2}$$

where $\mathbf{F}_k$ ($N_k \times Q$) denotes the component score matrix of the k-th group, $\mathbf{B}$ ($J \times Q$) denotes the loading matrix which is identical for all groups and therefore does not have an index k, and $\mathbf{E}_k$ ($N_k \times J$) denotes the matrix of residuals. In SCA-P, the most general variant, the variances of the component scores over all groups are fixed at one. This restriction only partly identifies the solution, in that the components of an SCA solution can be freely rotated without altering the fit of the solution. In SCA-P, the variances of and the correlations between the retrieved components may vary across the groups. Consequently, it may occur that a specific component has little variance within particular groups, or that two components have a very high correlation for one group and almost no correlation for the other groups. Apart from that, SCA-P leaves no room to find differences in covariance structure between groups.

To more extensively trace between-group differences and similarities in the component structure, clusterwise SCA (De Roover et al., 2012) was developed. Clusterwise SCA partitions the K groups into C clusters and models the data of the groups within each cluster with a simultaneous component model. In this paper, we will use the most general clusterwise SCA variant, i.e., clusterwise SCA-P (De Roover et al., 2013b), which applies SCA-P (see above) within each cluster. In this paper, given that we assume that each group is characterized by the same latent factor structure (apart from differences in the factor variances and covariances), we restrict the number of components Q of the cluster-specific SCA-P models to be the same across the clusters (for other purposes, clusterwise SCA extensions exist that allow the numbers of components to vary across clusters; see De Roover et al., 2013a).

Formally, clusterwise SCA-P models the data of one group as follows:

$$\mathbf{X}_k = \sum_{c=1}^{C} p_{kc}\mathbf{F}_k\mathbf{B}^{(c)'} + \mathbf{E}_k \tag{3}$$

where p_{kc} denotes the entries of the binary partition matrix $\mathbf{P}$ ($K \times C$) which equal 1 when group k is assigned to cluster c and 0 otherwise and $\mathbf{B}^{(c)}$ ($J \times Q$) is the loading matrix of cluster c ($c = 1, \ldots, C$). Given that the SCA-P models per cluster are independent of one another, the cluster-specific components can be freely rotated within each cluster.

To fit a clusterwise SCA-P solution with C clusters and Q components to a given data set, the sum of the squared residuals is minimized by means of an alternating least squares (ALS) algorithm (more details can be found in De Roover et al., 2013b). A multistart procedure is used to reduce the probability of ending up in a local minimum.

MODEL SELECTION

When applying clusterwise SCA-P analysis, the number of clusters C and components Q need to be specified by the user. In the context of measurement invariance analysis, the number of components Q is equal to the number of latent variables under study, but the most appropriate number of clusters is usually unknown. To deal with this model selection problem, clusterwise SCA-P solutions are estimated using 1 to $C^{\max}$ clusters. Next, a scree test (Cattell, 1966) is performed to determine the number of clusters after which the increase in fit levels off: C^{best}. Specifically, C^{best}

[4]One may argue that this advice is somewhat inconsistent with the fact that measurement invariance testing is usually done on raw data. Multigroup CFA is less sensitive to between-item differences in variability, however, due to the many restrictions on the factor loadings (i.e., zero or equality constraints). Moreover, if the variance differences are relatively small, analyzing unrescaled data (i.e., data that is only within-population centered) will yield very similar results to those obtained when overall rescaling is performed. For instance, for our illustrative data set (see Application section), the clustering of the groups was identical and the subset of non-invariant items consisted of the items reported below with the exception of "proud about myself." The CFI indices of the multigroup CFA models that are obtained when these six items are removed, are almost identical to the reported ones.

is the C-value that maximizes the following scree ratio $sr_{(C)}$ (see also Ceulemans and Kiers, 2006, 2009):

$$sr_{(C)} = \frac{\text{VAF}_C - \text{VAF}_{C-1}}{\text{VAF}_{C+1} - \text{VAF}_C} \quad (4)$$

where VAF_C is the percentage of variance-accounted-for of a solution with C clusters (and Q components; for software to perform the scree test; see Wilderjans et al., 2013). VAF_C is calculated as the fitted sum of squares divided by the total sum of squares:

$$\text{VAF}_C = 100 \times \frac{\sum_{k=1}^{K}\sum_{c=1}^{C} p_{kc} \left\| \mathbf{F}_k \mathbf{B}^{(c)} \right\|^2}{\sum_{k=1}^{K} \|\mathbf{X}_k\|^2} \quad (5)$$

Of course, differences in VAF_C-values may be very small when the data contain only a few non-invariant items. Therefore, when in doubt about the optimal number of clusters, it is advised to perform the detection procedure (see below) using different C-values to examine the stability of the obtained set of non-invariant items, taking into account that the higher the C-value, the larger the number of non-invariant items may become.

DETECTION OF NON-INVARIANT ITEMS

To detect non-invariant items, we propose to apply the following procedure[5], which consists of four steps:

1. Rotate cluster-specific loadings toward the postulated factor structure: Since clusterwise SCA-P solutions have rotational freedom (see above), the comparability of the cluster-specific component loadings is optimized by orthogonally rotating them toward a target matrix that corresponds to the factor model specification that was used in the measurement invariance testing (taking loadings equal to one if an item is assumed to load on a factor and zero otherwise).
2. Screen for the presence of non-invariant items: Calculate, for each cluster pair and for $q = 1, \ldots, Q$, the Tucker's congruence coefficient φ (Tucker, 1951) between the qth cluster-specific components. The congruence coefficient is an index of similarity between components (or factors). It takes values between -1 and 1, where a negative value indicates that one of the components should be reflected, a value of zero indicates no agreement, a value between 0.85 and 0.95 indicates high similarity, and a value higher than 0.95 corresponds to virtual identity (Lorenzo-Seva and ten Berge, 2006). Therefore, in what follows, we will assume that components are identical if the congruence value is 0.96 or larger. Next, the minimal φ-value $\varphi_{\min}$ across these $C(C-1)/2 \times Q$ congruence coefficients is calculated. When $\varphi_{\min}$ is less than 0.96, this suggests that the data contain non-invariant items and the procedure continues. When $\varphi_{\min}$ is 0.96 or larger, there is no indication that non-invariant items are present. Thus, the procedure is stopped and it is concluded that the clusterwise SCA-P analysis endorses weak measurement invariance. Note that the congruence coefficient measures the proportionality of two sets of component loadings and is thus insensitive to differences in component scale (which influence the loading sizes due to the restrictions on the component variances).
3. Detect which items are non-invariant: Remove each item one by one (i.e., with replacement) from the loading matrices and recompute the minimum congruence coefficient $\varphi_{\min}$ (across all cluster pairs and components), re-rotating the remaining loadings toward the corresponding subset of the target matrix. The item for which the absolute value of this $\varphi_{\min}$ is the highest (which indicates that the between-cluster congruence of the components improves the most when omitting this item) is considered non-invariant and permanently removed. This step is repeated until the resulting $\varphi_{\min}$ value exceeds 0.96, indicating weak invariance.
4. Re-estimate the cluster-specific SCA-P models for the remaining subset of items and repeat steps 1–3 to check whether additional non-invariant items are found. Continue until no more non-invariant items seem to be present (i.e., $\varphi_{\min} > 0.96$). Note that the clustering is fixed in this step. Allowing an update of the clustering would often lead to a different, nonsensical clustering, because the removal of non-invariant items diminishes the differences driving the initial clustering.

This procedure differs in three important respects from the CFA procedures that were discussed in the introduction: firstly, our procedure examines the non-invariance of complete items, whereas the sequential model modification procedure and item-level invariance testing focus on the non-invariance of each loading separately. Secondly, whereas the CFA tests examine either configural or weak invariance, the procedure proposed above captures both simultaneously. Thirdly, clusterwise SCA is more parsimonious than the three CFA procedures in that it examines differences between clusters of groups rather than between separate groups, which possibly lowers the capitalization on chance.

APPLICATIONS

DATA DESCRIPTION

In this section, we will illustrate our method for detecting non-invariant items by means of data that were originally collected to investigate emotional acculturation. Emotional acculturation refers to the process by which immigrants' patterns of emotional experience assimilate to those of the host culture (De Leersnyder et al., 2011). To investigate the robustness of the phenomenon, the researchers examined two different host cultures, and included minority groups from different heritage cultures (the cultures from which the immigrants stem). Moreover, to compare the emotional patterns of the immigrants with those of their heritage culture, two heritage groups were inspected as well (see **Table 1** for an overview of the groups involved).

First, as previous research found emotional differences between independent and interdependent cultural contexts (e.g., Mesquita, 2001; Kitayama et al., 2006), the host and heritage cultures under study differ along the independent-interdependent dimension, with both host cultures (European American and Belgian contexts) on the independent end and all heritage cultures

[5]The procedure as well as the clusterwise SCA-P analyses are implemented in a Matlab R2013b function which can be obtained freely from the first author.

Table 1 | The 13 cultural groups under consideration and associated host country, design and sample size (note: each situation-subject combination counts as one observation).

Cultural group	Host country	Design	Removed observations due to missing data	Retained observations (N_i)	Partition
European Americans 1	USA	1	12	120	1
Korean immigrants	USA	1	21	126	1
Mexican immigrants	USA	1	16	188	1
East-Asian immigrants	USA	2	5	159	1
Latino immigrants	USA	2	1	142	1
European Americans 2	USA	2	10	122	1
Koreans	Korea	2	22	298	1
Flemish students 1	Belgium	3	5	183	2
Flemish students 2	Belgium	3	20	516	2
Belgian community	Belgium	3	26	166	2
Turkish 2nd generation immigrants	Belgium	3	17	157	2
Turkish 1st generation immigrants	Belgium	3	22	143	3
Turkish students	Turkey	3	119	699	3

The last column indicates to which cluster the cultural group is assigned in the clusterwise SCA-P model with three clusters and two components per cluster.

(Korea/East Asia, Mexico/Latino, and Turkey) on the interdependent end. A second reason for focusing on these host and heritage cultures is that they differ considerably from an acculturation point of view. The US and Belgian cultural contexts have different migration histories that translate in different policies and different collective ideas on immigrants and immigration (Van Acker, 2012). Within the US context, Korean/East Asian minorities differ from Mexican/Latino minorities in terms of both education and employment; the former are highly educated, and work white collar jobs, whereas the latter are typically less educated, and occupy blue collar jobs. Within the Belgian context, Turkish minorities tend to have little education and occupy more working class (as opposed to middle class) jobs than majority members. One of the Belgian majority samples was matched with respect to education and socio-economic status to the Turkish minority sample; the other two Belgian majority samples consisted of Belgian (Flemish) university students.

The participants reported on one to four specific situations that differed on the dimensions of valence (positive, negative), social engagement (socially engaged, socially disengaged), and social context (with friends, at home/with family, at school/work). They then rated on a 7-point Likert scale to what extent they experienced each of 17 different emotions (see **Table 3**). The situations were chosen according to three types of design. In Design 1, participants received three emotional prompts that pertained to the same type of emotional situation (e.g., positive disengaging situation), but that differed with respect to social context. In Design 2, participants received four emotional prompts that pertained to the same social context (e.g., family), but that differed with respect to type of emotional situation (i.e., positive disengaging situation, positive engaging situation, negative disengaging situation, negative engaging situation). Design 3 was similar to Design 2, but due to time constraints, participants only completed two types of emotional prompts for the same social context. The design was fixed within each group (see **Table 1**), which implies that differences between cultural groups may have been confounded with differences in design. Note that we removed observations (i.e., subject-situation combinations) with missing data from the data set (see **Table 1**).

Of course, the fact that the data contain up to four observations per subject may introduce some dependencies among the observations within a group, violating the independence assumption of the CFA framework. Retaining only one observation per subject would drastically reduce the sample size per group, leading to convergence problems when performing (multigroup) CFA analyses. However, given that for the majority of the subjects only one or two observations are included in the data (i.e., 289 subjects with one observation and 819 subjects with two observations) and that varying the type and context of the emotional situations causes substantial within-subject differences, we deem the degree of dependence in the data to be limited and not prohibitive for using the current data as an illustration for our proposed procedure.

The questionnaires (i.e., the prompts) were developed in English and then translated from English into Korean, Spanish, Dutch and Turkish, and then back-translated into English by bilingual researchers. In this pragmatic type of translation (Brislin, 1980), the accuracy of meaning is emphasized, rather than a literal, word-for-word translation.

CONFIGURAL AND WEAK INVARIANCE TESTING

A latent variable structure that seems reasonable for this data set is one with a positive emotions factor and a negative emotions one (Kuppens et al., 2006). Therefore, we tested the configural and weak invariance of this latent variable structure by means of the R packages Lavaan 0.5–15 (Rosseel, 2012) and SemTools 0.4–0). To take the ordinal nature of the Likert scale ratings into account, we used the diagonally weighted least squares (DWLS) estimator (Jöreskog and Sörbom, 1996, pp. 23–24). **Table 2** contains the comparative fit indices (CFI) for the CFA model for each group separately, as well as for a multigroup CFA model without imposing further equality restrictions (to evaluate configural invariance) and a multigroup CFA model with equal loadings for all groups (to evaluate weak invariance). We focused on the CFI

Table 2 | Comparative fit indices (CFI) for multigroup CFA analyses imposing positive affect and negative affect factors for the emotional acculturation data.

	All 17 emotions	Seven non-invariant emotions removed
GROUP-SPECIFIC FIT		
European Americans 1	**0.91**	0.99
Korean immigrants	**0.87**	0.97
Mexican immigrants	**0.81**	**0.90**
East-Asian immigrants	**0.93**	1.00
Latino immigrants	**0.89**	0.97
European Americans 2	0.97	1.00
Koreans	0.96	0.99
Flemish students 1	0.97	0.99
Flemish students 2	0.95	0.99
Belgian community	**0.94**	0.99
Turkish 2nd generation immigrants	0.97	1.00
Turkish 1st generation immigrants	0.98	1.00
Turkish students	0.97	0.98
OVERALL FIT		
Multigroup CFA	0.95	0.98
Multigroup CFA with equal loadings across groups	**0.86**	0.96

CFI values lower than 0.95 are in bold face.

because it is a fit index that also performs well in small samples (Hu and Bentler, 1999), which is an advantage considering the small sample size for some of the cultural groups. A CFI value of 0.95 suggests a good fit of the model to the data (Hu and Bentler, 1999), a CFI between 0.90 and 0.95 corresponds to a reasonable fit (Bentler, 1990; Tabachnick and Fidell, 2005), and a CFI value lower than 0.90 indicates a bad fit (Bentler, 1990).

First, we examined configural invariance by looking at the CFI value of the unconstrained multigroup model. The CFI value is 0.95; thus, at first sight the baseline model with the positive and negative affect factors seemed to be appropriate (i.e., configural invariance confirmed). However, the CFI values for the separate groups conveyed that this model had an excellent fit for some groups but not for all, with CFI < 0.90 for the Korean, Mexican, and Latino immigrants.

Second, we looked at the overall fit of the weak invariance model (i.e., equal loadings across all groups). The CFI value of 0.86, and also the difference of 0.09 in CFI with the overall configural invariance model, indicated a bad fit of the model to the data and, thus, a flat out rejection of weak invariance.

CLUSTERWISE SCA AND THE DETECTION OF NON-INVARIANT ITEMS

To investigate whether the lack of invariance is due to the presence of non-invariant items, we centered the data per group and normalized them over groups and applied clusterwise SCA-P analyses with 1–6 clusters and two components per cluster. A scree plot with the VAF values of the resulting models is presented in **Figure 1**. Although fit differences are small, the increase in fit clearly levels off after three clusters. This is also confirmed by the scree ratio's, which amount to 1.9, 2.3, 1.3, and 1.1 for two, three, four, and five clusters, respectively. Thus, we proceeded with the clusterwise SCA-P model with three clusters and two components per cluster.

The corresponding partition of the cultural groups is shown in **Table 1**. The cultural groups living in the USA are gathered in Cluster 1, together with the Koreans. Cluster 2 consists of the indigenous Belgian groups, together with the second generation Turkish immigrants in Belgium. Cluster 3 contains the Turkish students living in Turkey and the first generation Turkish immigrants in Belgium. The fact that the second generation Turkish immigrants were assigned to the Belgian cluster suggests that these immigrants acculturated with respect to the meaning of their emotions. The assignment of the Korean immigrants as well as the Koreans to the USA cluster, indicates that—in case of three clusters and two components—neither of them stands out enough in terms of their covariance structure to end up in a separate cluster. Note that all the data in cluster 1 were gathered by means of Designs 1 and 2, whereas the data in clusters 2 and 3 were collected using Design 3 only. Thus, it is possible to clearly interpret the differences between cluster 2 and 3 as differences in cultural groups, whereas the differences between cluster 1 on the one hand and clusters 2 and 3 on the other, may be due to design differences, in addition to cultural differences.

The target (i.e., positive emotions component and negative emotions component) rotated loadings of the three clusters are given in **Table 3**. At first sight, the component structure in all three clusters closely resembles this target structure: a first component that mainly corresponds to the positive emotions and a second component that is mainly constituted by negative emotions. Similarity to the target structure was corroborated by the congruence values between the cluster-specific components and the corresponding columns of the target structure, which always exceeded 0.85 indicating high similarity—but not identity—to the target structure (see **Table 4**).

However, we did notice some remarkable between-cluster differences for specific items. For instance, "surprised" has a high loading on the "positive" component in the Turkish cluster and a moderately high positive loading on the "negative" component in the USA cluster. These differences were confirmed by the Tucker's congruence coefficients between the corresponding cluster-specific components (see **Table 4**), which lay between 0.90 and 0.95, indicating between-cluster differences in loading structure.

Applying the procedure described in the Methods section, yielded the following seven non-invariant items: "strong," "proud about myself," "surprised," "relying," "resigned," "bored," and "indebted." After removing the seven non-invariant items and estimating a new SCA-P model per cluster for the retained subset of variables, the congruence coefficients of the components between clusters ranged from 0.96 to 0.99. Against the background of other research on the cultures of comparison, it is possible to meaningfully interpret some of these non-invariant items. For instance, "proud about myself" has a higher negative loading on the "negative" component in the Belgian cluster. This indicates

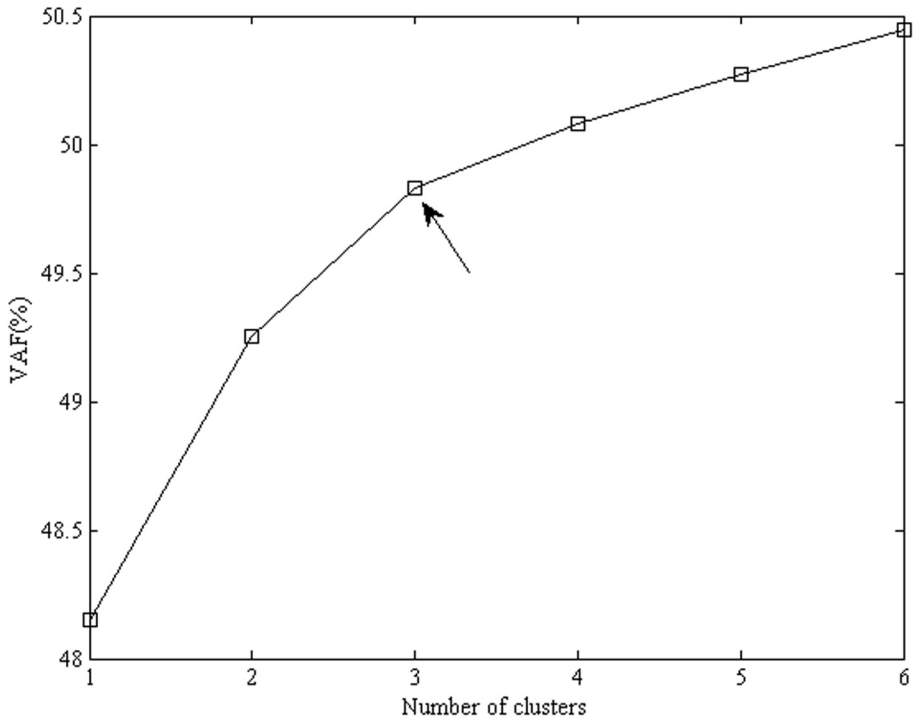


FIGURE 1 | Percentage of explained variance for clusterwise SCA-P solutions for the emotional acculturation data, with the number of components equal to 2 and the number of clusters varying from 1 to 6. The favored number of clusters is 3 (indicated by the arrow), because the increase in fit levels off after three clusters.

that when Belgians experience negative emotions, they feel less proud about themselves than people belonging to the other cultural groups. The association between negative emotions and feeling less proud is also, to a lesser extent, observed in the USA and Koreans cluster. The different meaning of "proud about myself" between the Turkish cluster on the one hand, and the Belgian and USA and Koreans cluster may be understood in the light of the specific meaning that this concept takes on in cultures that emphasize "independence" (e.g., Markus and Kitayama, 1991): in these cultures, pride has the connotation of being successful and superior (Roseman, 2013), and thus may be seen as compromised by failure which is associated with negative emotions.

As another example, "relying" has a moderately high positive loading on the "negative" component in the USA and Koreans cluster. Follow-up analyses showed that the negative connotation of "relying" in this cluster is mainly driven by the clear negative connotation among the European Americans (in an SCA-P model for the two groups of European Americans "relying" had a loading of 0.42 on the negative component), which is less outspoken in the USA immigrant groups (loading of 0.27) and among Korean natives (loading of 0.24). The feeling of relying on someone else may have a negative connotation (and co-occur with negative emotions) for the European Americans, because it clashes with central ideals of personal autonomy and self-reliance (e.g., Markus and Kitayama, 1991).

Another interesting difference is the fact that "resigned" has a lower loading on the "negative" component in the Belgian and Turkish clusters in comparison to the USA and Koreans cluster. Moreover, in the Turkish cluster, "resigned" loads primarily on the "positive" component. The different meanings of "resigned" may be associated with different ideas on control. Personal control is a central value in middle class American culture, where it is considered instrumental to an individual's independence and autonomy (Markus and Kitayama, 1991); in this context, resignation is likely to have the negative connotation of giving up. On the other end of the spectrum, Turkish culture emphasizes "kismet" or fate: Turkish people tend to have a strong belief in both fate (e.g., Ergüder et al., 1991) and authority (Dağ, 1991; Lester et al., 1991). Therefore, feeling resigned may be regarded as positive in the Turkish culture, as it denotes that one accepts an event and one's fate.

To summarize, important differences in component structure were found, indicating that a subset of the emotions covary differently with the other emotions or are even valued differently in some of the cultural groups. Surely, these cross-cultural differences are interesting in itself. Furthermore, these differences may be what's hampering the measurement invariance testing, as they pertain to both the primary (e.g., "surprised" being less strongly associated with the "positive" component in the USA and Turkish clusters) and secondary loadings (e.g., "resigned" being part of positive affect in the Turkish cluster), which may, respectively, explain the rejection of the weak invariance model and the bad fit of the configural invariance model for some of the groups.

Table 3 | Cluster-specific loadings for the clusterwise SCA-P model with three clusters and two components per cluster, orthogonally Procrustes rotated toward a positive and negative target structure.

Emotions	Cluster 1 (USA and Koreans)		Cluster 2 (Belgian)		Cluster 3 (Turkish)	
	Pos.	Neg.	Pos.	Neg.	Pos.	Neg.
Respect	**0.76**	−0.18	**0.72**	−0.20	**0.86**	−0.19
Interested	**0.69**	−0.24	**0.63**	−0.26	**0.69**	−0.11
Helpful	**0.75**	−0.17	**0.63**	−0.12	**0.61**	−0.14
Close	**0.64**	−0.23	**0.74**	−0.02	**0.79**	−0.24
Strong	***0.66***	*−0.28*	***0.47***	***−0.46***	***0.75***	*−0.28*
Proud about myself	***0.64***	***−0.43***	***0.49***	***−0.58***	***0.68***	*−0.34*
Relying	***0.53***	*0.30*	***0.76***	*−0.01*	***0.78***	*−0.09*
Surprised	*0.26*	*0.30*	*0.31*	*−0.12*	***0.72***	*−0.13*
Ill feelings	−0.35	**0.51**	−0.39	**0.55**	−0.39	**0.69**
Upset	−0.35	**0.74**	−0.37	**0.62**	**−0.47**	**0.69**
Irritated	−0.25	**0.69**	**−0.57**	**0.54**	−0.20	**0.67**
Embarrassed	−0.12	**0.79**	−0.18	**0.60**	−0.18	**0.73**
Ashamed	−0.09	**0.81**	−0.19	**0.77**	−0.16	**0.65**
Guilty	−0.22	**0.73**	−0.14	**0.82**	−0.26	**0.69**
Bored	*0.07*	***0.40***	*−0.25*	***0.35***	***−0.51***	***0.74***
Indebted	***0.45***	***0.42***	*0.27*	***0.74***	*0.27*	***0.53***
Resigned	*−0.08*	***0.60***	*0.06*	*0.32*	***0.40***	*0.33*

Loadings larger than 0.40 in absolute value are indicated in bold face. Non-invariant items are indicated in italic.

Table 4 | Tucker's congruence coefficients between the cluster-specific component loadings in Table 3 and the target structure (per component), as well as between the cluster-specific components mutually (per component and per cluster pair), when including all variables.

		Cluster 2		Cluster 3		Target structure	
		Positive	Negative	Positive	Negative	Positive	Negative
Cluster 1	**Positive**	0.94	-	0.90	-	0.89	-
	Negative	-	0.93	-	0.93	-	0.90
Cluster 2	**Positive**			0.94	-	0.86	-
	Negative			-	0.95	-	0.88
Cluster 3	**Positive**					0.89	-
	Negative					-	0.94

MODIFIED CONFIGURAL AND WEAK INVARIANCE TESTING

To examine whether the detected non-invariant items were indeed contributing to the violations of measurement invariance, we removed these seven items from the data and re-evaluated the configural and weak invariance CFA models mentioned above. Regarding configural invariance, the resulting CFI values for the separate groups were higher (see **Table 2**), leaving only one group (i.e., Mexican immigrants) in the reasonable fit range (i.e., CFI between 0.90 and 0.95) and none in the bad fit range (i.e., CFI $<$ 0.90). The same holds for the CFI value of the multigroup model, which amounted to 0.98 and suggested an excellent overall fit. Regarding weak measurement invariance, the corresponding CFA model had a good CFI of 0.96, as compared to 0.86 when all items were included. Surely, the fit decrease of 0.02 when going from the configural invariance model to the weak invariance model (i.e., from 0.98 to 0.96) was still large enough to reject weak invariance, but clearly our procedure pinpointed some interesting differences in emotion covariances that were interfering with weak invariance.

Another strategy for incorporating the results of our procedure in the CFA testing is freeing some of the loadings of the non-invariant items. Regarding configural invariance, we added secondary loadings (instead of zero ones) for the non-invariant items. The overall CFI for the resulting multigroup CFA model was 0.98, whereas the group-specific fit values were very similar to those in **Table 2**. Regarding weak invariance, allowing both loadings of the non-invariant items to vary across groups yielded a partial weak invariance model with a CFI value of 0.96.

RESULTS OF CFA METHODS FOR DEALING WITH INVARIANCE VIOLATIONS

To compare our results to those of popular CFA methods for dealing with invariance violations, we applied the three procedures discussed in the Introduction. In the sequential modification procedure (MacCallum et al., 1992; Stuive et al., 2009), we confined ourselves to modifying the weak invariance model by allowing primary loadings to differ in certain groups or adding secondary loadings for certain groups, because several authors have reported that this modification procedure outperforms methods which allow for other modifications (e.g., including residual covariances; MacCallum, 1986; Silvia and MacCallum, 1988). We continued freeing or adding loadings for specific groups, as specified by the modification indices, until the resulting increase in fit (ΔCFI) no longer exceeded 0.01. As a result, the primary loading of "bored" was freed for group 4 and a secondary and free loading was added for "resigned," also for group 4. The CFI of the resulting partial weak invariance model is 0.86.

The item-level invariance testing (Cheung and Rensvold, 1999) entailed no less than 66 additional invariance tests (see Introduction); i.e., two factor-specific tests, 28 tests for the non-zero loadings on the positive factor and 36 tests for the non-zero loadings on the negative factor. The integrated results of these tests indicate that the primary loadings of "surprised," "relying," "resigned," "bored," and "helpful" have to be freed across the groups. The CFI of the thus obtained partial weak invariance model is 0.92.

The strategy presented by Byrne and van de Vijver (2010) involved two times 17 additional multigroup CFA analyses; i.e., deleting one item at a time, for configural invariance on the one hand and for weak invariance on the other hand. With respect to configural invariance, only one item yielded a CFI increase of more than 0.01 upon deletion: "indebted." Thus, for "indebted," there seemed to be some misfit with respect to the imposed factor structure, possibly due to the need for a secondary loading of indebted on the positive component for some of the groups. Deleting "indebted" led to an overall CFI of 0.97 and group-specific fit values ranging from 0.85 to 0.99 with only the Mexican

immigrants having a CFI below 0.90 (i.e., 0.85, implying bad fit). With respect to items "interested", "helpful", "close", "relying", "ill feelings", "embarrassed", and "ashamed", no decision could be made, since the corresponding multigroup CFA analyses (i.e., with one of these items being deleted) did not converge. With respect to weak invariance, five non-invariant items were traced by this approach: "surprised," "relying," "resigned," "bored," and "indebted." When deleting this subset of items the overall CFI of the multigroup CFA with equal loadings across groups amounted to 0.96.

CONCLUSION WITH RESPECT TO THE CROSS-CULTURAL EMOTION DATA

This application demonstrated that using clusterwise SCA to investigate what is causing a lack of measurement invariance makes sense, because (1) the fit of the multigroup CFA models improved greatly when the detected non-invariant items were removed or when the models were modified by allowing for specific secondary loadings or by letting particular primary loadings vary across the groups, and (2) the detected set of non-invariant items largely overlapped with those resulting from the three multigroup CFA procedures. Also, the unique aspect of the proposed approach—the clustering of the groups—was nicely illustrated, i.e., meaningful clusters of groups were found and the non-invariant items could be traced by comparing the loadings between these clusters, without having to inspect the loadings of each group separately. Moreover, the total CPU time of the clusterwise SCA-P analyses, i.e., including the model selection and the detection procedure was about 33 s only (using Matlab R2013b on an Intel® Core™ i7-3770K processor of a personal computer, with a clock frequency of 3.4–3.9 GHz and a RAM speed of 1600 MHz) while the item-level invariance testing and the Byrne and van de Vijver (2010) approach were much more cumbersome and time-consuming (on the same computer, the former procedure took more than 24 h to run and the latter about 2 h and a half, using the R-packages Lavaan 0.5–15 and SemTools 0.4–0). Applying the sequential model modification procedure took only 8 min, but this was because it led to only two modifications with a ΔCFI > 0.01 (and, consequently, did not improve the model fit very much).

GENERAL DISCUSSION

The issue of measurement invariance is ubiquitous in the behavioral sciences nowadays as more and more studies yield multivariate multigroup data. Although CFA based methods have been proposed to trace which items are hampering measurement invariance, these methods have some disadvantages in that they require researchers to run a multitude of analyses and in that they imply assumptions that are often questionable. In this paper, we proposed an alternative strategy which consists of running clusterwise SCA and comparing the resulting loadings via congruence coefficients to quickly trace possible non-invariant items. The cross-cultural application demonstrated that this novel approach is useful and can co-exist with the traditional CFA approaches.

As also holds for the discussed CFA approaches, it may sometimes occur that invariance is still rejected after removing the items indicated as non-invariant by the new approach. In such cases, one may consider the following actions to further pursue invariance.

Firstly, it may be that the number of clusters was too small to detect all non-invariant items. Thus, it may be useful to examine a clusterwise SCA solution with more clusters—for the complete set of items—and repeat the detection heuristic.

Secondly, when the overall fit of the baseline multigroup CFA model is still bad, this suggests that the CFA model is misspecified. For example, additional factors may be needed to approach a good fit, the postulated latent variable model may be completely off or distributional assumptions may be violated. If so, the clusterwise SCA based detection approach will not be able to remedy this problem and neither can the CFA approaches. To get more grip on what is going on, exploratory factor analysis may be used to examine the factor structure. Moreover, problems with regard to the target structure can be easily traced from the clusterwise SCA results by checking whether the congruence coefficients between the cluster-specific components and the postulated factors are low.

Thirdly, when the fit of the baseline CFA model remains below standards for only one or a few of the groups after removing the detected non-invariant items, it may be that the group(s) in question need other CFA model modifications such as residual covariances. To this end, one may resort to the group-specific modification indices.

Fourthly, when configural invariance is established but weak invariance is still rejected, a more strict congruence criterion may be needed for the data at hand (e.g., 0.97 instead of 0.96) to detect all subtle size differences in loadings which may be causing the rejection of weak invariance. Especially when the number of invariant items is much larger than the number of non-invariant items, it may happen that the congruence criterion is not strict enough to detect the most subtle differences.

Fifthly, it may be the case that the factor structure is appropriate for most groups but incorrect for a minority of outlying groups. Clusterwise SCA will conveniently assign these outlying groups to one or more separate clusters, with the congruence coefficients between the corresponding cluster-specific and the a priori factor structure being low. For such data, one may want to remove the outlying groups and repeat the measurement invariance testing. In this regard, Byrne and van de Vijver (2010) specified a set of criteria to identify groups that are possibly outlying in terms of their item scores and evaluated the goodness-of-fit of the multigroup CFA model when deleting these groups one by one (i.e., with replacement). However, these criteria are based on the level of the items[6] rather than on their factor structure. This implies that this approach is not ideal to track groups with outlying factor structures.

Finally, it may be that measurement invariance simply cannot be established because the groups form a few clusters that are characterized by a distinct factor structure. Using clusterwise SCA, one can conveniently discern such clusters and perform the measurement invariance testing within the clusters. Since, up to

[6]Note that a convenient approach for identifying countries that are deviant with respect to item level was proposed by Ceulemans et al. (2013). This approach was based on robust principal component analysis.

now, no factor analytic counterpart exists, clusterwise SCA is the only method to find clusters of groups based on within-group component or factor structure without having to resort to tedious pairwise comparisons of group-specific structures.

As a final remark, an attractive feature of the proposed approach is that its applicability, unlike the CFA based methods, largely surpasses the context of measurement invariance. Indeed, the approach can also be used when researchers do not have an a priori idea about the underlying structure of the items and about possible differences across groups (e.g., Krysinska et al., in press). To this end, a standard SCA-P analysis (i.e., without clustering) is run on the data and the resulting component loadings—the "common component structure"—are used as the target structure toward which the cluster-specific loadings are rotated.

AUTHOR NOTES

Kim De Roover is a post-doctoral fellow of the Fund for Scientific Research Flanders (Belgium). The research leading to the results reported in this paper was sponsored in part by Belgian Federal Science Policy within the framework of the Interuniversity Attraction Poles program (IAP/P7/06), and by the Research Council of KU Leuven (GOA/2010/02).

REFERENCES

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238

Brannick, M. T. (1995). Critical comments on applying covariance structure modelling. *J. Organ. Behav.* 16, 201–213. doi: 10.1002/job.4030160303

Brislin, R. W. (1980). "Translation and content analysis of oral and written materials," in *Handbook of Cross-Cultural Psychology: Vol. 2. Methodology*, eds H. C. Triandis and J. W. Berry (Boston, MA: Allyn and Bacon), 137–164.

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

Byrne, B. M., and van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: addressing the issue of nonequivalence. *Int. J. Testing* 10, 107–132. doi: 10.1080/15305051003637306

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behav. Res.* 1, 245–276. doi: 10.1207/s15327906mbr0102_10

Ceulemans, E., Hubert, M., and Rousseeuw, P. (2013). Robust multilevel simultaneous component analysis. *Chemometr. Intell. Lab. Syst.* 129, 33–39. doi: 10.1016/j.chemolab.2013.06.016

Ceulemans, E., and Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: a numerical convex hull based method. *Br. J. Math. Stat. Psychol.* 59, 133–150. doi: 10.1348/000711005X64817

Ceulemans, E., and Kiers, H. A. L. (2009). Discriminating between strong and weak structures in three-mode principal component analysis. *Br. J. Math. Stat. Psychol.* 62, 601–620. doi: 10.1348/000711008X369474

Cheung, G. W., and Rensvold, R. B. (1999). Testing factorial invariance across groups: a reconceptualization and proposed new method. *J. Manage.* 25, 1–27. doi: 10.1177/014920639902500101

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equ. Model.* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Dağ, I. (1991). Rotter'in iç-diÕ kontrol oda.i ölçe.inin üniversite ö.rencileri için güvenirli.i ve geçerli.i [The reliability and validity of Rotter's Internal-External Locus of Control Scale for university students]. *Turk. Psychol. Assoc. J.* 7, 10–16.

De Leersnyder, J., Mesquita, B., and Kim, H. S. (2011). Where do my emotions belong? A study of immigrants' emotional acculturation. *Pers. Soc. Psychol. Bull.* 37, 451–463. doi: 10.1177/0146167211399103

De Roover, K., Ceulemans, E., Timmerman, M. E., Nezlek, J. B., and Onghena, P. (2013a). Modeling differences in the dimensionality of multiblock data by means of clusterwise simultaneous component analysis. *Psychometrika* 78, 648–668. doi: 10.1007/s11336-013-9318-4

De Roover, K., Ceulemans, E., Timmerman, M. E., and Onghena, P. (2013b). A clusterwise simultaneous component method for capturing within-cluster differences in component variances and correlations. *Br. J. Math. Stat Psychol.* 86, 81–102. doi: 10.1111/j.2044-8317.2012.02040.x

De Roover, K., Ceulemans, E., Timmerman, M. E., Vansteelandt, K., Stouten, J., and Onghena, P. (2012). Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychol. Methods* 17, 100–119. doi: 10.1037/a0025385

Ergüder, Ü., Esmer, Y., and Kalaycioğlu, E. (1991). *Türk Toplumunun Degáerleri [Values in Turkish culture]*. Istanbul: Tüsiad Yayinlari, Tüsiad, publication number T/91, 6.145.

Gorsuch, R. (1983). *Factor Analysis, 2nd Edn.* Hillsdale, NJ: L. Erlbaum Associates.

Gorsuch, R. L. (1990). Common factor analysis versus component analysis: some well and little known facts. *Multivariate Behav. Res.* 25, 33–39. doi: 10.1207/s15327906mbr2501_3

Green, B. F. (1976). On the factor score controversy. *Psychometrika* 41, 263–266. doi: 10.1007/BF02291843

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychol. Methods* 6, 430–450. doi: 10.1037/1082-989X.6.4.430

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika* 36, 409–426. doi: 10.1007/BF02291366

Jöreskog, K. G., and Sörbom, D. (1996). *LISREL 8: User's Reference Guide, 2nd Edn.* Chicago, IL: Scientific Software International.

Kelloway, E. K. (1995). Structural equation modelling in perspective. *J. Organ. Behav.* 16, 215–224. doi: 10.1002/job.4030160304

Kiers, H. A. L., and ten Berge, J. M. F. (1994). Hierarchical relations between methods for simultaneous components analysis and a technique for rotation to a simple simultaneous structure. *Br. J. Math. Stat. Psychol.* 47, 109–126. doi: 10.1111/j.2044-8317.1994.tb01027.x

Kitayama, S., Mesquita, B., and Karasawa, M. (2006). The emotional basis of independent and interdependent selves: socially disengaging and engaging emotions in the US and Japan. *J. Pers. Soc. Psychol.* 91, 890–903. doi: 10.1037/0022-3514.91.5.890

Krysinska, K., De Roover, K., Bouwens, J., Ceulemans, E., Corveleyn, J., Dezutter, J., et al. (in press). Measuring religious attitudes in (post-)secularised Western European context: recent changes in the underlying dimensions of the Post-Critical Belief Scale. *Int. J. Psychol. Relig.* doi: 10.1080/10508619.2013.879429

Kuppens, P., Ceulemans, E., Timmerman, M. E., Diener, E., and Kim-Prieto, C. H. U. (2006). Universal intracultural and intercultural dimensions of the recalled frequency of emotional experience. *J. Cross Cult. Psychol.* 37, 491–515. doi: 10.1177/0022022106290474

Lawley, D. N., and Maxwell, A. E. (1962). Factor analysis as a statistical method. *Statistician* 12, 209–229. doi: 10.2307/2986915

Lester, D., Castromayor, I. J., and Içli, T. (1991). Locus of control, depression, and suicidal ideation among American, Philippine, and Turkish students. *J. Soc. Psychol.* 13, 447–449. doi: 10.1080/00224545.1991.9713873

Lorenzo-Seva, U., and ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology* 2, 57–64. doi: 10.1027/1614-2241.2.2.57

MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychol. Bull.* 100, 107–120. doi: 10.1037/0033-2909.100.1.107

MacCallum, R. C., Roznowski, M., and Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychol. Bull.* 111, 490–504. doi: 10.1037/0033-2909.111.3.490

Markus, H. R., and Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychol. Rev.* 98, 224–253. doi: 10.1037/0033-295X.98.2.224

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Meredith, W., and Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Med. Care*, 44, S69–S77. doi: 10.1097/01.mlr.0000245438.73837.89

Mesquita, B. (2001). Emotions in collectivist and individualist contexts. *J. Pers. Soc. Psychol.* 80, 68–74. doi: 10.1037/0022-3514.80.1.68

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

Muthén, B., and Asparouhov, T. (2013). *BSEM Measurement Invariance Analysis (MplusWeb Notes No.17)*. Available online at: https://www.statmodel.com/examples/webnotes/webnote17.pdf (Accessed March 24, 2014).

Roseman, I. J. (2013). Appraisal in the emotion system: coherence in strategies for coping. *Emot. Rev*. 5, 141–149. doi: 10.1177/1754073912469591

Rosseel, Y. (2012). Lavaan: an R package for structural equation modeling. *J. Stat. Softw.*, 48, 1–36.

Silvia, E. S. M., and MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behav. Res.* 23, 297–326. doi: 10.1207/s15327906mbr2303_2

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *Br. J. Math. Stat. Psychol*. 27, 229–239. doi: 10.1111/j.2044-8317.1974.tb00543.x

Steiger, J. H. (1989). *EzPATH: A Supplementary Module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT.

Stuive, I., Kiers, H. A. L., and Timmerman, M. E. (2009). Comparison of methods for adjusting incorrect assignments of items to subtests: oblique multiple group method versus confirmatory common factor method. *Educ. Psychol. Measur.* 69, 948–965. doi: 10.1177/0013164409332226

Tabachnick, B. G. and Fidell, L. S. (2005). *Using Multivariate Statistics, 5th Edn*. Boston, MA: Pearson.

Timmerman, M. E., and Kiers, H. A. L. (2003). Four simultaneous component models of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika* 86, 105–122. doi: 10.1007/BF02296656

Tucker, L., R. (1951). *A Method for Synthesis of Factor Analysis Studies (Personnel Research Section Rep. No. 984)*. Washington, DC: Department of the Army.

Van Acker, K. (2012). *Flanders' Real and Present Threat: How Representations of Intergroup Relations Shape Attitudes Towards Muslim Minorities.* Doctoral dissertation, University of Leuven, Belgium. ISBN: 978-94-6190-938-1

Velicer, W. F., and Jackson, D. N. (1990). Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. *Multivariate Behav. Res.* 25, 1–28. doi: 10.1207/s15327906mbr2501_1

Wilderjans, T. F., Ceulemans, E., and Meers, K. (2013). CHull: a generic convex hull based model selection method. *Behav. Res. Methods* 45, 1–15. doi: 10.3758/s13428-012-0238-5

Williams, R., and Thomson, E. (1986). Normalization issues in latent variable modeling. *Sociol. Methods Res*. 15, 24–43. doi: 10.1177/0049124186015001002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 January 2014; accepted: 29 May 2014; published online: 20 June 2014.
Citation: De Roover K, Timmerman ME, De Leersnyder J, Mesquita B and Ceulemans E (2014) What's hampering measurement invariance: detecting non-invariant items using clusterwise simultaneous component analysis. Front. Psychol. ***5****:604. doi: 10.3389/fpsyg.2014.00604*
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

METHODS ARTICLE
published: 30 October 2014
doi: 10.3389/fpsyg.2014.01216

Testing for measurement invariance and latent mean differences across methods: interesting incremental information from multitrait-multimethod studies

Christian Geiser[1]*, G. Leonard Burns[2] and Mateu Servera[3]

[1] *Department of Psychology, Utah State University, Logan, UT, USA*
[2] *Department of Psychology, Washington State University, Pullman, WA, USA*
[3] *Department of Psychology, University Research Institute of Health Sciences (IUNICS), University of Balearic Islands, Palma, Spain*

Edited by:
Rens Van De Schoot, Utrecht University, Netherlands

Reviewed by:
Moritz Heene, Ludwig Maximilian University of Munich, Germany
Marjolein Verhoeven, Utrecht University, Netherlands

****Correspondence:***
Christian Geiser, Department of Psychology, Utah State University, 2810 Old Main Hill, Logan, UT 84322-2810, USA
e-mail: christian.geiser@usu.edu

Models of confirmatory factor analysis (CFA) are frequently applied to examine the convergent validity of scores obtained from multiple raters or methods in so-called multitrait-multimethod (MTMM) investigations. We show that interesting incremental information about method effects can be gained from including mean structures and tests of MI across methods in MTMM models. We present a modeling framework for testing MI in the first step of a CFA-MTMM analysis. We also discuss the relevance of MI in the context of four more complex CFA-MTMM models with method factors. We focus on three recently developed multiple-indicator CFA-MTMM models for structurally different methods [the correlated traits-correlated (methods – 1), latent difference, and latent means models; Geiser et al., 2014a; Pohl and Steyer, 2010; Pohl et al., 2008] and one model for interchangeable methods (Eid et al., 2008). We demonstrate that some of these models require or imply MI by definition for a proper interpretation of trait or method factors, whereas others do not, and explain why MI may or may not be required in each model. We show that in the model for interchangeable methods, testing for MI is critical for determining whether methods can truly be seen as interchangeable. We illustrate the theoretical issues in an empirical application to an MTMM study of attention deficit and hyperactivity disorder (ADHD) with mother, father, and teacher ratings as methods.

Keywords: multitrait-multimethod (MTMM) analysis, measurement invariance, measurement equivalence, mean and covariance structures, mean differences across raters, random vs. fixed methods, rater agreement

Multitrait-multimethod (MTMM) analysis is frequently used to examine the convergent and discriminant validity of psychological measurements based on measurement designs in which multiple constructs or traits are assessed by multiple methods (Campbell and Fiske, 1959; Widaman, 1985; Millsap, 1995; Dumenci, 2000). In the classical MTMM design, multiple (typically at least three) methods are used to assess multiple (typically at least three) constructs or traits. The analysis of MTMM data has historically focused on the interpretation of the so-called *MTMM matrix*, which summarizes the correlations between variables in an MTMM design. The MTMM matrix approach was developed by Campbell and Fiske (1959) who also proposed heuristics for the interpretation of MTMM correlations in terms of convergent and discriminant validity. Over the years, confirmatory factor analysis (CFA) has become a popular tool for analyzing data obtained from MTMM designs, given the greater flexibility of the CFA framework compared to the original MTMM matrix approach (for a detailed discussion of the advantages of the CFA approach to MTMM analyses, see Eid et al., 2006).

Whereas Campbell and Fiske's original approach focused exclusively on correlation structures, CFA models allow analyzing not only correlation, but also covariance and mean structures (e.g., Little, 1997). Moreover, CFA models allow for the analysis of multiple (instead of just a single) indicators (e.g., items or scales) per trait-method unit (TMU). For example, self-, parent, and teacher-reports on three or more items or scales could be used to assess depression. Using multiple indicators per TMU has the advantage that researchers can study the factorial validity at the item level for each type of method, that method effects can be analyzed separately for different traits to examine the potential trait-specificity of method effects, and that measurement error influences (unreliability) can be more properly estimated (Marsh and Hocevar, 1988; Eid et al., 2003). Despite the fact that modern CFA methods allow for an analysis of covariance *and* mean structures in the same model, most applied MTMM studies so far have focused exclusively on modeling covariance structures. In addition, most MTMM studies still use single-indicator designs (i.e., just a single observed variable per TMU; e.g., Servera et al., 2010).

In the present article, we show that by moving from an exclusively covariance- or correlation-based MTMM approach to an approach that includes latent means, more fine-grained information about convergent validity and method effects can be gained in CFA-MTMM analyses. In this context, we highlight a specific advantage of multiple-indicator MTMM designs that has received little attention in the MTMM literature so far: the possibility to test for measurement invariance (MI) across multiple raters or methods when the different methods provided scores on comparable measurement instruments (e.g., equivalent questionnaires).

Analyzing mean structures in MTMM models and the investigation of mean method effects in MTMM models has been proposed in previous work (Eid, 2000; Pohl et al., 2008; Pohl and Steyer, 2010). The new aspect in the current paper is the investigation of MI across methods, which facilitates a proper interpretation of mean method effects. Examining MI is a novel aspect in MTMM research and considering MI itself as well as for the interpretation of mean method effects adds important information when evaluating MTMM data.

Although some MTMM studies have examined MI in the context of multiple-group comparisons (i.e., for comparing measurement structures across different populations; e.g., Cole and Maxwell, 1985; Marsh et al., 1992), the issue of MI across methods within the same population seems to have received little attention in the literature. For example, although Woehr et al. (2005) tested for configural, metric, and residual invariance across different raters, they did not examine intercept invariance or latent mean differences across raters.

In the present paper, we focus on MTMM designs that (1) use multiple raters as methods and (2) equivalent questionnaires across raters. Such designs are common in the applied MTMM literature. For example, Cole et al. (1997) used equivalent child and parent versions of questionnaires measuring depression and anxiety in children. Similarly, Grigorenko et al. (2010) used self-report, parent-report, and teacher-report versions of the same questionnaire to assess problem behaviors in children. Burns et al. (in press) assessed symptoms of hyperactivity, impulsivity, inattention, and academic impairment in 5th graders by mother, father, and teacher ratings, all of which filled out equivalent forms of a questionnaire.

We show that by studying MI across raters, additional information about method effects can be obtained that cannot be revealed through purely correlational MTMM analyses. By testing for MI, researchers can first of all examine whether the same factor structure holds across methods (configural invariance, see discussion below)—an assumption that is often implicitly made in MTMM studies, but rarely formally tested. In addition, researchers can examine whether different methods (e.g., different raters) use the questionnaire scales in a similar way (i.e., whether the scales have equal difficulty and discrimination across raters in the sense of item response theory). For example, when the same symptoms of attention deficit and hyperactivity disorder (ADHD) are rated by parents and teachers, different loadings or intercepts may be obtained, showing that the observed symptom scores differ in difficulty or discrimination between raters. This could, for example, indicate that teachers are more lenient than parents in their ratings or that certain symptoms are only weakly related to the latent variable for a specific type of rater. Therefore, the finding of measurement non-invariance across raters can reveal additional insights into more subtle forms of method effects.

In addition to the general relevance of MI testing across methods, researchers may be uncertain as to the relevance of MI in different CFA-MTMM models with method factors. With the present article, we also want to contribute to a better understanding of the issue of MI in the context of recently developed CFA-MTMM models. In line with modern MTMM approaches, we focus on models that use multiple indicators per TMU (Marsh and Hocevar, 1988; Eid et al., 2003, 2008; Geiser et al., 2012).

We first explain why the inclusion of means in addition to covariances and testing for MI can reveal useful incremental information in MTMM studies in general. We then present a modeling framework for testing MI in MTMM studies that use multiple indicators per TMU. Subsequently, we turn to four different models with method factors that have recently been proposed for the analysis of MTMM data. For each of the four models, we discuss which level of MI these models require for a proper interpretation of the model parameters.

ANALYZING MEAN STRUCTURES IN MTMM ANALYSES: INTERESTING INCREMENTAL INFORMATION

The reported outcome of most MTMM studies are statistical indices that provide information on the convergent validity (or consistency) of different methods or raters in terms of the rank order of the individuals that were assessed by the different methods. As a simple example, researchers often interpret a correlation between, say parent and teacher ratings of child behavior in terms of convergent validity, following Campbell and Fiske's (1959) guidelines. In terms of individual differences, such a correlation coefficient indicates to which extent different raters agree as to the rank order of children on the outcome variable (e.g., depression, externalizing problem behavior, ADHD). This information is clearly useful, as it informs us about how much variability is shared between raters or methods for the same construct.

Here, we argue that covariance-based information on multiple raters' agreement as to the relative standing of individuals on a construct (which is typically the focus of MTMM studies) is not the only useful information that can be gained from MTMM studies. This is because information about the overall level (mean) is usually also of interest. That is, we argue that researchers often want to know, for instance, whether parent ratings of problem behaviors result in the same or similar conclusions about the overall level of these behaviors in a population as do teacher ratings. Such questions can be addressed by analyzing mean structures in CFA-MTMM models in addition to covariance structures, which is a relatively novel aspect in MTMM research. Comparing means across raters requires a certain level of MI across raters. That is, for such comparisons to be meaningful, the measurement parameters that link the observed scores to the latent variables should be equal across raters to ensure comparable scales. This issue parallels the comparison of latent means across groups in multigroup CFA and structural equation modeling (SEM) as well as the examination of mean changes across time in longitudinal studies (e.g., Little, 1997).

THE MEANING OF MI FOR MTMM DATA

Formally, MI can be said to hold in an MTMM study if (1) a similar factor structure is found for different methods used to assess multiple traits or constructs using multiple indicators per TMU and/or (2) certain parameters of the measurement model (e.g., factor loadings, intercepts, or residual variances) that relates the observed scores to latent variables are equal across methods. Condition (1) requires only that (a) the same number of factors be found across methods and (b) the pattern of loadings (which variable loads onto which factor) be the same across methods. For Condition 1, the term *configural invariance* has been coined in the general MI literature (e.g., Meredith, 1993; Widaman and Reise, 1997; Millsap, 2011). Condition 2 is more restrictive and requires that not only the basic factor structure be equivalent across methods, but also specific parameters such as factor loadings, intercepts, or residual variances.

Even though it seems clear that establishing at least configural invariance (equal factor structure) across methods is a necessity for a meaningful comparison across methods, even configural invariance is typically not formally tested in MTMM studies (for exceptions, see Woehr et al., 2005; Burns et al., in press). Here, we argue that testing for MI is useful when different methods were scored on comparable scales (e.g., multiple raters taking the same questionnaire), because such analyses (1) provide additional insights into method effects and (2) allow researchers to test whether it is meaningful to compare latent means across raters. A meaningful comparison of latent means across methods requires that at least strong MI be established across methods (i.e., equal loadings and intercepts). Strong MI ensures that the origin and units of measurement are the same across raters.

A MODELING FRAMEWORK FOR TESTING MI IN MTMM STUDIES

Marsh and Hocevar (1988) proposed a general target or baseline model for MTMM studies that use multiple indicators per TMU. In the present article, we show that this model as well as an extension of it can be used for testing MI across methods in MTMM studies. Marsh and Hocevar's model is depicted in **Figure 1** as a path diagram. For simplicity, here we consider only a single construct (or trait; $j = 1$) that is measured by just two methods ($k = 1, 2$). Each TMU jk is represented by three indicators ($i = 1, 2, 3$). Focusing on this simple design is sufficient to explain the general MI issues, which can then easily be generalized to larger MTMM designs. (In our empirical application presented later on, we used a design with one trait and three methods).

Note that in our path diagrams, we represent both the covariance and mean structure, following the RAM conventions introduced by McArdle (1980). The model proposed by Marsh and Hocevar (1988) includes a separate common factor (or true score variable) for each TMU (e.g., one factor for mother ratings of hyperactivity and one factor for teacher ratings of the same construct). All TMU factors are allowed to correlate. Marsh and Hocevar's model has a number of advantages for MTMM analyses in general. First, the model allows testing the appropriateness of the latent factor structure for each TMU. For example, the assumption of unidimensionality may be violated for some or all methods, thus providing evidence against configural invariance across methods, which is fundamental to MTMM analysis. Second, the model allows examining Campbell and Fiske's (1959) MTMM correlations at the latent level. That is, rather than inspecting observed correlations that are attenuated by measurement error as in Campbell and Fiske's original approach, the model in **Figure 1** provides the same correlations at the level of common true score variables. Therefore, the MTMM correlations are corrected for random measurement error. This has the advantage that the estimated correlations are less biased and easier to compare between constructs with different scale reliabilities[1].

In the present article, we focus on the possibility to formally test for MI across methods within each construct or trait by estimating constrained versions of the model. In these constrained versions, parameters of the measurement model such as loadings, intercepts, or residual variances are constrained to be equal across methods to test whether and to which extent such MI assumptions are tenable. In order to test for MI, we can examine the following series of models in line with Widaman and Reise (1997)[2]:

1. A model of configural invariance postulates the same factor structure (number of factors and pattern of loadings) across methods, but does not impose any formal equality constraints on non-zero factor loadings, intercepts, or residual variances.
2. A model of weak invariance postulates the same factor structure plus equal factor loadings for corresponding indicators across methods.
3. A model of strong invariance postulates the same factor structure, equal factor loadings, and equal intercepts for corresponding indicators across methods.
4. A model of strict invariance postulates the same factor structure, equal factor loadings, equal intercepts, and equal measurement error (residual) variances for corresponding indicators across methods.

One drawback of the model presented in **Figure 1** is that it will only fit MTMM data when the indicators are strictly homogeneous in the sense that within each method, all indicators have perfectly correlated true score variables that differ only in scaling (i.e., have potentially different intercepts and loadings). This assumption is frequently violated in practice, because different items or subscales often measure slightly different facets of a construct and therefore do not share exactly the same common true score variable in the sense of classical test theory models. For example, one item for measuring the construct depression may refer to *sadness*, whereas another item meant to measure the same construct may refer to *sleeping problems*. Therefore, although both

[1]Latent correlations are less biased than observed correlations only if correlated errors of measurement do not exist or if such correlated errors are properly modeled.

[2]In addition to full invariance that requires all corresponding parameters to be equal across raters, models of partial invariance have also been discussed in the general MI literature. Partial MI means that invariance is tenable only for a subset, but not all indicators (e.g., Byrne et al., 1989). In addition, in the MTMM case, measurement parameters may be invariant across some, but not all methods as shown in the empirical example section.

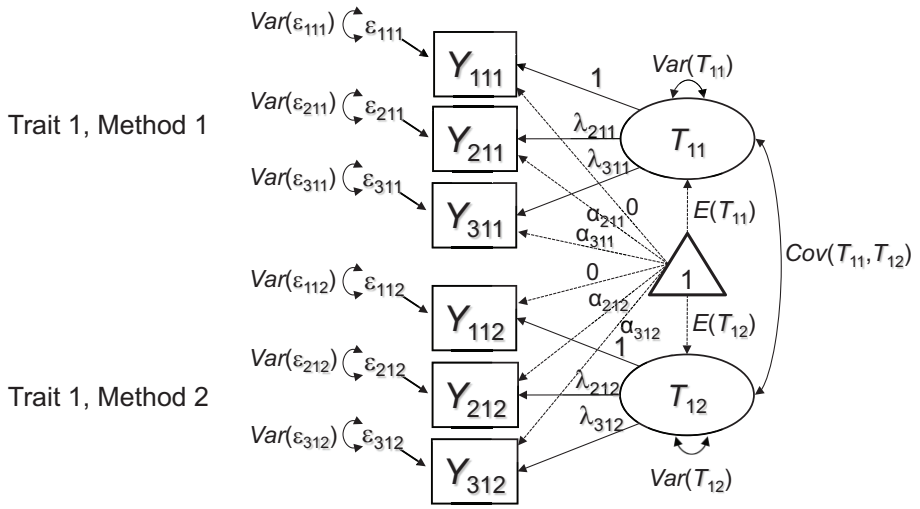


FIGURE 1 | CFA measurement model for multiple-indicator MTMM data. Each latent factor T_{jk} represents the error-free (true) scores of a specific TMU. The picture shows an example in which three indicators Y_{ijk} ($i = 1, 2, 3$) are used to measure one construct or trait ($j = 1$) by two methods ($k = 1, 2$).

items measure facets of depression, they may not share exactly the same true score variable. As another example, consider item wording effects due to positive and negative item wording (e.g., Vautier et al., 2003), which can also cause a common true score model to show misfit.

If such inhomogeneities generalize across methods (e.g., if parent ratings of sadness are more strongly correlated with teacher ratings of sadness than with teacher ratings of sleeping problems), then the model in **Figure 1** likely will not fit the data very well, because this model assumes a homogeneous correlation structure across methods for the same construct. We therefore present an extension of the model in **Figure 1**, in which this issue is addressed by including indicator-specific residual factors for all but a reference indicator (see **Figure 2**). An equivalent approach has been presented previously to account for indicator heterogeneity in longitudinal studies, in which the same issues occur when the same indicators are repeatedly measured across time in single-method designs (Eid et al., 1999).

The model in **Figure 2** uses a reference-indicator approach in which $I - 1$ (of a total number of I) indicators are contrasted against a reference indicator (without loss of generality, the first indicator $i = 1$ in **Figure 2** is chosen as reference indicator). This is done by introducing residual method (or indicator-specific) factors IS_{ij} for all except the reference indicator. These indicator-specific factors have means of zero and are by definition uncorrelated with the true score that represents the reference indicator (see Appendix A in Supplementary Material for the formal definition of these factors). The IS_{ij} factors reflect indicator-specific variance that is not shared with the reference indicator, but is shared across methods. Indicator-specific factors can be correlated with each other in principle, reflecting potential shared deviations of non-reference indicators from the reference indicator (e.g., the reference indicator measuring the sadness aspect of depression, whereas the remaining two indicators both refer to sleeping problems). Whether or not these correlations are meaningful and should be estimated depends on the specific application.

The model with indicator-specific factors can be used for MI testing in the same way as the model in **Figure 1**. If at least strong MI (i.e., equal reference factor loadings λ_{ijk} and intercepts α_{ijk}) can be established in either the model in **Figure 1** or the model in **Figure 2**, latent mean differences across methods can be meaningfully interpreted[3]. In the model with indicator-specific factors, it is also possible to test for invariant loadings γ_{ijk} of the IS_{ij} factors (in addition to the reference factor loadings λ_{ijk}).

Establishing invariant IS_{ij} factor loadings is not necessary for a meaningful comparison of the reference factors T_{1jk} across methods (for more detailed explanations see Appendix A in Supplementary Material). In many applications, only the factors T_{1jk} and their invariance across methods will be of substantive interest. Nonetheless, comparisons of the γ_{ijk} loadings across methods can reveal interesting information about the extent to which indicator-specific effects are reflected in different methods. For example, some methods may not be as sensitive to subtle differences in item content as others. This can be reflected in non-invariant γ_{ijk} loadings across methods. In the following section, we examine the issue of MI in the context of more sophisticated CFA-MTMM models with method factors that researchers often use in a second step of an MTMM analyses. Subsequently, we present applications of all models to an actual data set.

[3]Some authors have recommended that strict invariance be established before latent mean comparisons are conducted (see, e.g., Wu et al., 2007). Strict invariance is only necessary, however, when correlated errors of measurement exist and are not properly modeled. We recommend that researchers pay careful attention to tests of model fit to detect potential error correlations and that such correlations—if they exist—be properly modeled with additional latent variables.

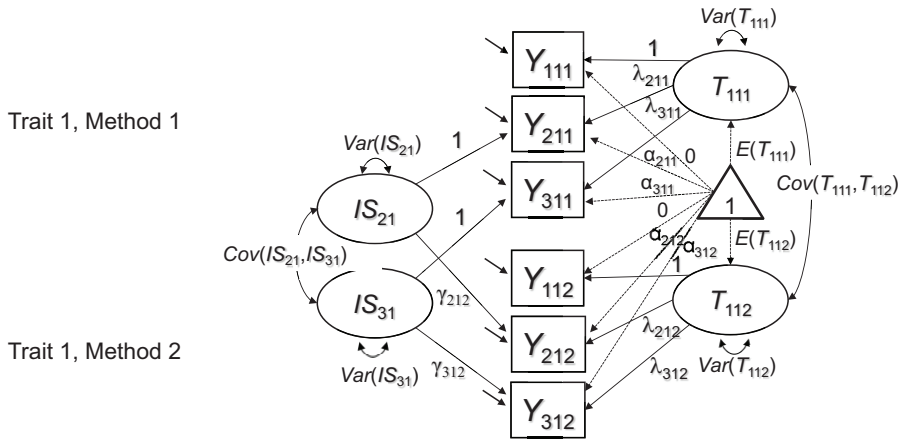


FIGURE 2 | Extended CFA measurement model for multiple-indicator MTMM data. In contrast to **Figure 1**, the extended model contains $I - 1$ indicator-specific factors IS_{ij} to reflect shared indicator-specific effects across raters. The latent factors T_{1jk} are now specific to the reference indicator Y_{1jk} and therefore carry an additional index for the reference indicator.

DIFFERENT MTMM MODELS

More sophisticated CFA-MTMM models are often employed because the simple CFA models in **Figures 1**, **2** do not directly express method effects in terms of latent variables (i.e., method factors), except for indicator-specific effects. In contrast, more complex CFA-MTMM models contain additional latent variables that directly reflect method effects in terms of latent methods factors. Such models allow explicitly contrasting different methods against a gold standard method (e.g., Eid, 2000; Pohl et al., 2008) or against a common trait (Pohl and Steyer, 2010). Furthermore, more complex models allow relating method effects to external variables, which is not possible in Marsh and Hocevar's (1988) simple CFA model discussed above.

In this article, we focus on four CFA-MTMM models that are relatively new: (1) Eid et al.'s (2003) multiple indicator CT-C(M – 1) model, (2) Pohl et al.'s (2008) latent difference model, (3) Pohl and Steyer's (2010) latent means model, and (4) Eid et al.'s (2008) CFA-MTMM model for interchangeable methods. Whereas the first three models were developed for use with structurally different methods (e.g., different fixed types of raters such as mothers, fathers, and teachers, which are not drawn from the same set of raters), Eid et al.'s (2008) CFA-MTMM model was developed for interchangeable (random) methods (e.g., randomly selected customers rating a product or service). Furthermore, whereas the first three models can all be defined as equivalent versions of Marsh and Hocevar's (1988) simple CFA model, the CFA-MTMM model for interchangeable raters in general implies a different covariance and mean structure.

We focus on the above models, because all of them can be formulated based on the well-defined concepts of classical testy theory (CTT). This ensures that in all of the models, the trait and method factors have a clear meaning and interpretation (Geiser et al., 2014a). Given the fact that the CT-C(M – 1), latent difference, and latent means approaches each imply the exact same measurement model as the simple CFA model presented previously, we show only the structural parts of the models for simplicity and parsimony in **Figure 3**.

THE CT-C(M – 1) APPROACH

Presentation of the model

Figure 3A shows the structural part of the CT-C($M - 1$) model in the version first presented by Geiser et al. (2008) and discussed in detail in Geiser et al. (2012)[4]. In the CT-C($M - 1$) model, one method serves as gold standard or reference method. This could either be a method that a researcher has most confidence in or that is most different from the remaining methods (for guidelines as to the choice of the reference method, see Geiser et al., 2008, 2012). For example, Geiser et al. (2014b) examined the convergent validity of giftedness assessments in children using the CT-C($M - 1$) approach. They selected a maximum-performance test battery to serve as reference method, given that the test battery provided a more objective measure of abilities relative to more subjective ability ratings provided by the children themselves, their parents, and their teachers.

In **Figure 3A**, without loss of generality, the first method ($k = 1$) was chosen as reference. The second and any additional methods are regressed on the true score variable pertaining to the reference method using a latent regression analysis:

$$E\left(T_{1jk}|T_{1j1}\right) = \beta_{0jk} + \beta_{1jk}T_{1j1}$$

[4]The CT-C(M – 1) model version is slightly more restricted than Eid et al.'s (2003) original model version. We chose to present the more restricted version here, given its direct correspondence to and mathematical equivalence with the latent difference and latent means models. The differences between Eid et al.'s (2003) model and the version presented here are explained in Geiser et al. (2012); the general MI issues discussed below apply to either version of the model.

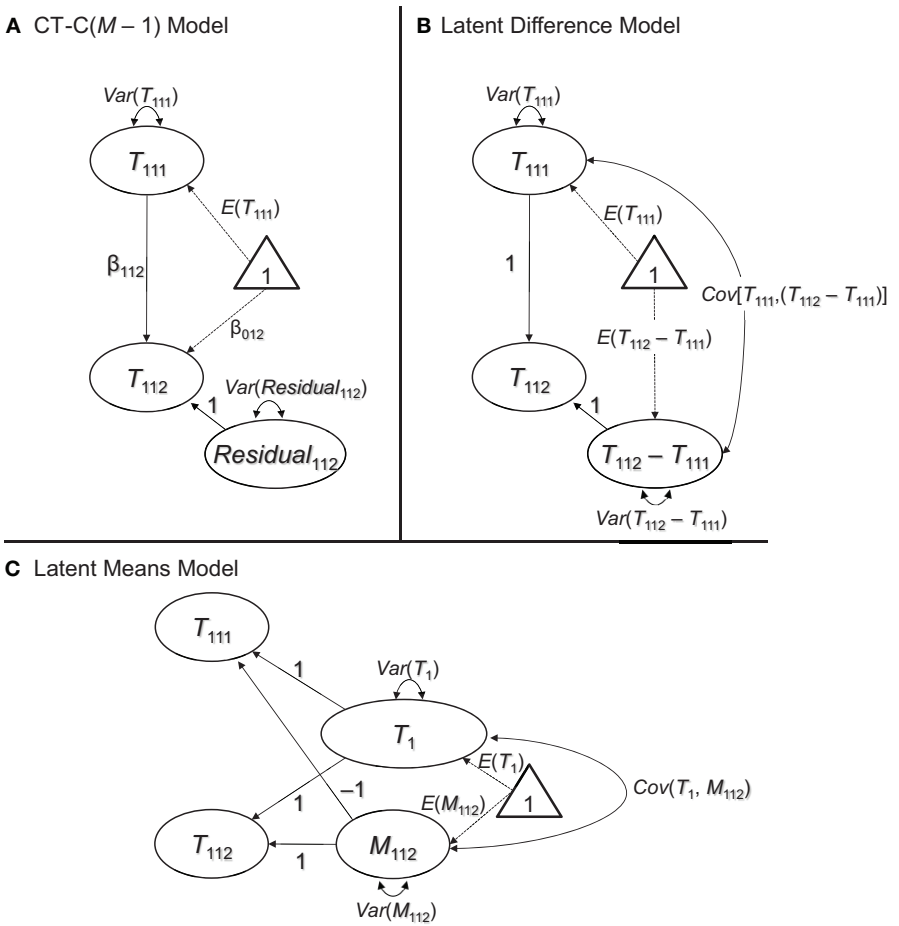


FIGURE 3 | Three different ways to examine method effects with structurally different methods. (A) Latent regression [CT-C(M – 1)] approach. **(B)** Latent difference approach. **(C)** Latent means approach. All three approaches imply the same covariance and mean structure at the latent level, but differ in terms of which level of MI they require (see discussion in the text). The measurement part of the models is the same as in **Figures 1**, **2** and therefore not shown in this figure.

where β_{0jk} and β_{1jk} indicate regression coefficients and $k \neq 1$. The residuals of these regressions

$$\text{Residual}_{1jk} = T_{1jk} - E\left(T_{1jk} |\ T_{1j1}\right)$$

serve as method factors in the CT-C($M-1$) model. Note that the CT-C($M-1$) model has the same number of parameters in the structural model as the simple CFA model. For one construct and two methods, there are five structural parameters: the reference factor mean and variance, the regression coefficients β_{0jk} and β_{1jk}, and the latent residual [method factor] variance. Note that for more than two traits or methods, admissible covariances among latent factors would be additional parameters to be estimated in the structural model.

Even though mathematically equivalent, the CT-C($M-1$) model represents a useful extension of the simple CFA model, because it allows us to express the information about method effects (defined relative to a reference method) in terms of latent method factors that are residuals with respect to the reference factors. Given that the method factors are defined as residuals relative to the reference factors, they are by definition uncorrelated with the reference factors and thus represent independent variance components (Eid et al., 2003). It also follows from their definition as residuals that the method factors have means of zero. Hence, it would not be meaningful to make statements about method factor means in the CT-C($M-1$) model.

The CT-C($M-1$) model allows us to (a) quantify what percentage of the observed or true score variance in different methods is shared vs. not shared with the reference method and (b) directly relate method effects to other variables (e.g., by correlating method factors with external variables). The proportion of observed variance that is shared with the reference method is expressed by the consistency coefficient:

$$Con\left(Y_{ijk}\right) = \lambda_{ijk}^2 \beta_{1jk}^2 Var\left(T_{1j1}\right) \ / \ Var\left(Y_{ijk}\right).$$

The consistency coefficient is often used as an indicator of convergent validity relative to the reference method or "gold standard." The proportion of observed variance that is not shared

with the reference method is expressed by the method-specificity coefficient:

$$MSpe\left(Y_{ijk}\right) = \lambda_{ijk}^{2} Var\left(Residual_{1jk}\right) \; / \; Var\left(Y_{ijk}\right).$$

The method-specificity coefficient is used to indicate which portion of the observed variance is unique to a specific method and not shared with the reference method. Correlations between method factors are allowed in the CT-C(M – 1) model. These correlations are partial correlations between non-reference methods from which variance shared with the reference method has been partialled out. Therefore, method factor correlations reflect a shared perspective (or "bias") of non-reference methods relative to the reference method.

MI in the CT-C(M – 1) model

The CT-C(M – 1) model allows contrasting different methods against a reference method by means of a latent regression approach. For this purpose, strictly speaking, MI across methods beyond configural invariance is not required. That is, for the interpretation of the *standardized* regression coefficients as well as the coefficients of consistency and method-specificity, it does not matter whether different methods were measured on the same scale, because the coefficients of interest are standardized. This makes the CT-C(M – 1) model very flexible for examining the convergent validity of different methods. For example, Geiser et al. (2014b) examined the convergent validity of objective ability tests and subjective ability ratings. Objective and subjective assessments were made on completely different scales; nonetheless the CT-C(M – 1) model allowed examining the degree of convergent validity across these methods.

On the other hand, the interpretation of the *unstandardized* regression coefficients β_{0jk} and β_{1jk} can in some cases be difficult if the different methods used different scales. This issue parallels the potential difficulty of interpreting unstandardized regression coefficients in standard ordinary least squares regression analysis when predictor and criterion variables used different or arbitrary metrics. Furthermore, if a researcher wants to make comparisons of latent means across methods based on the latent mean of the reference factor and the unstandardized regression coefficients, strong MI is required in the same way as in Marsh and Hocevar's (1988) model.

THE LATENT DIFFERENCE APPROACH

Presentation of the model

The latent difference approach is closely related to the CT-C(M – 1) approach in that different methods are contrasted against a reference method. However, in the latent difference approach, method effects are defined as simple deviations (differences) from a reference method true score variable rather than as regression residuals (Pohl et al., 2008). Latent difference factors ($T_{1jk} - T_{1j1}$) are introduced that reflect method effects in terms of the difference between a true score of a non-reference method and the true score pertaining to the reference method (see **Figure 3B**):

$$T_{1jk} = 1T_{1j1} + 1\left(T_{1jk} - T_{1j1}\right).$$

The latent difference model again has the same number of parameters in the structural model as Marsh and Hocevar's (1988) simple CFA model (five parameters in the case of two methods: the reference factor mean and variance, the latent difference factor mean and variance, and the covariance between reference and latent difference factor). In contrast to the CT-C(M – 1) model, a correlation between reference and method factor is allowed in the latent difference model, because the method factor is not defined as a regression residual with respect to the reference factor. Moreover, in contrast to the CT-C(M – 1) approach, the mean of the method factor can be estimated as well and reflects the latent mean difference between two methods. A more detailed comparison of the latent difference and CT-C(M – 1) models can be found in Geiser et al. (2012).

MI in the latent difference model

In the latent difference model, convergent validity is assessed in terms of the latent difference between true score variables pertaining to different methods. Smaller differences indicate greater convergent validity relative to the reference method. MI plays a more important role in the latent difference model than in the CT-C(M – 1) model. Given that method effects are defined in terms of difference scores between the true score variables pertaining to different methods, strong MI is critical for a meaningful interpretation of the structural model parameters in the latent difference model. When strong MI does not hold, the interpretation of the latent difference scores can become difficult, because a violation of strong MI indicates that the true score variables pertaining to different methods may not be measured with comparable origin or units of measurement. In this case, persons' individual difference scores as well as the mean and variance of the latent difference factor would be difficult to interpret.

THE LATENT MEANS APPROACH

Presentation of the model

In the latent means model, method effects are defined as deviations from an average across true score variables (Pohl and Steyer, 2010). In the first step, a common trait factor T_j is defined by averaging across the true score variables that reflect different TMUs. In our example with just one trait and just two methods, we obtain:

$$T_j := \left(T_{1j1} + T_{1j2}\right) \; / \; 2,$$

where the ":=" sign indicates a definition. The method factors are defined as deviations from the common trait:

$$M_{1j1} := T_{1j1} - T_j,$$
$$M_{1j2} := T_{1j2} - T_j.$$

Given their definition as deviations from the same average, the method factors sum up to zero (i.e., $M_{1j1} + M_{1j2} = 0$). Therefore, in the case of two methods, we obtain the following deterministic relationship between the two method factors:

$$M_{1j1} = -M_{1j2}.$$

It is thus sufficient to include only $M - 1$ method factors as in the CT-C($M - 1$) and latent difference models (i.e., the last method factor is fully determined by the implicit sum-to-zero constraint and therefore redundant). Here, without loss of generality, we dropped the first method factor so that we obtain the following structural model shown in **Figure 3C**:

$$T_{1j1} = T_j - M_{1j2},$$
$$T_{1j2} = T_j + M_{1j2}.$$

All trait and all method factors in the latent means model can be correlated. As in the CT-C($M - 1$) and latent difference models, for two TMUs, we obtain a structural model with five free parameters (the common trait factor mean and variance, the method factor mean and variance, and the covariance between the common trait and the method factor).

MI in the latent means model

The latent means model defines a common trait T_j as the average of true score variables T_{1jk} that pertain to the same construct j. Such an average is typically only meaningful when the true score variables are measured on the same scale. A similar argument applies to the interpretation of the method factors in the model: A deviation of a particular method from the grand average is only meaningful if all methods used the same scale. Moreover, as in the latent difference model, establishing at least strong MI is crucial for the interpretation of the model parameters in the latent means model. One difference between the latent difference and latent means models in this regard is that the latent difference model allows for partial MI, whereas the latent means model does not. That is, as long as at least one non-reference method shows MI relative to the reference method, the latent difference between the two can be meaningfully interpreted. In the latent means model, however, the common trait will typically only have a clear interpretation if *all* methods show at least strong MI.

THE CFA-MTMM MODEL FOR INTERCHANGEABLE METHODS

Presentation of the model

Eid et al. (2008) showed that measurement designs with interchangeable methods imply different measurement models for modeling trait and method effects than do designs with structurally different methods. This is because the underlying random experiment differs for designs with structurally different vs. interchangeable methods. Eid et al. (2008) presented a multilevel CFA approach for modeling interchangeable methods each of which used the same items to rate each trait. Nussbeck et al. (2009) showed that the same model can also be estimated within the single-level CFA framework. Here, we consider Nussbeck et al.'s single-level CFA approach (rather than the multilevel version) for two reasons: (1) The single-level version of the model is easier to compare to the previously described models for structurally different methods, and (2) the single-level CFA approach is more flexible in terms of explicitly testing assumptions of MI than is the multilevel approach (this parallels the issue of testing MI in longitudinal latent state-trait models in the single- vs. multilevel CFA framework as described in detail in Geiser et al., 2013).

When methods are interchangeable, they are considered randomly drawn from a set of equivalent methods (Eid et al., 2008). Eid et al. (2008) as well as Nussbeck et al. (2009) showed that this structure implies a CFA model with M uncorrelated method factors (i.e., a separate method factor for each interchangeable method; see **Figure 4**). As can be seen in **Figure 4**, for one trait and two methods, we obtain measurement models that have the same structure as longitudinal latent state-trait models (e.g., Geiser et al., 2013) for multiple indicators and that are similar to so-called bifactor models (e.g., Reise, 2012).

Figure 4A is a version of the model for homogeneous indicators all of which measure exactly the same common trait factor T_j. **Figure 4B** shows a model version with indicator-specific trait factors T_{ij}. Indicator specific traits are useful to capture inhomogeneities among indicators in a similar way as was done with indicator-specific factors in the previously discussed models for structurally different methods. The method factors M_{jk} are defined as residuals with respect to the trait factor(s). As a consequence, the trait factors are by definition uncorrelated with all M_{jk} pertaining to the same construct j, and all M_{jk} factors have means of zero by definition. Similar to the CT-C($M - 1$) model, method effects are defined as regression residuals. In contrast to the CT-C($M - 1$) model, however, the trait factors are common to all methods (rather than specific to a reference method), and the method factors are uncorrelated across methods. This makes sense, because of the interchangeable nature of the ratings. In contrast to structurally different methods, with interchangeable methods, there is no one method that is particularly outstanding or special (or seen as a gold standard). Therefore, it makes sense to include general trait factors and uncorrelated method factors for each method.

Note that the common factors and the method factors in the interchangeable model have a different meaning than the trait and method factors in the latent means model. The common trait in the latent means model is defined as an average of true scores, and the method factors are defined as differences from this average. In contrast, the common factor in the interchangeable model is not an average and the method factors are not differences from an average, but residuals with respect to a common factor.

MI in the CFA model for interchangeable methods

The interchangeable model is an interesting case with regard to the issue of MI, because it is less obvious whether or not MI across methods is required or should be established in the model. For the interpretation of the traits and method factors in the model, MI does not appear to be necessary, because the traits are not defined as an average of true scores, and the method factors are not defined as difference scores relative to a reference true score or an average of true scores as in the latent difference and latent means models. Nonetheless, testing for invariant loadings and intercepts across methods is critical in this model as well, albeit for different, somewhat more subtle reasons.

If different supposedly interchangeable methods result in different loadings or intercepts for the same indicator, this can question the assumption that these methods are truly interchangeable in the sense that they represent "random samples" drawn from a set of uniform methods. For example, a researcher may ask a

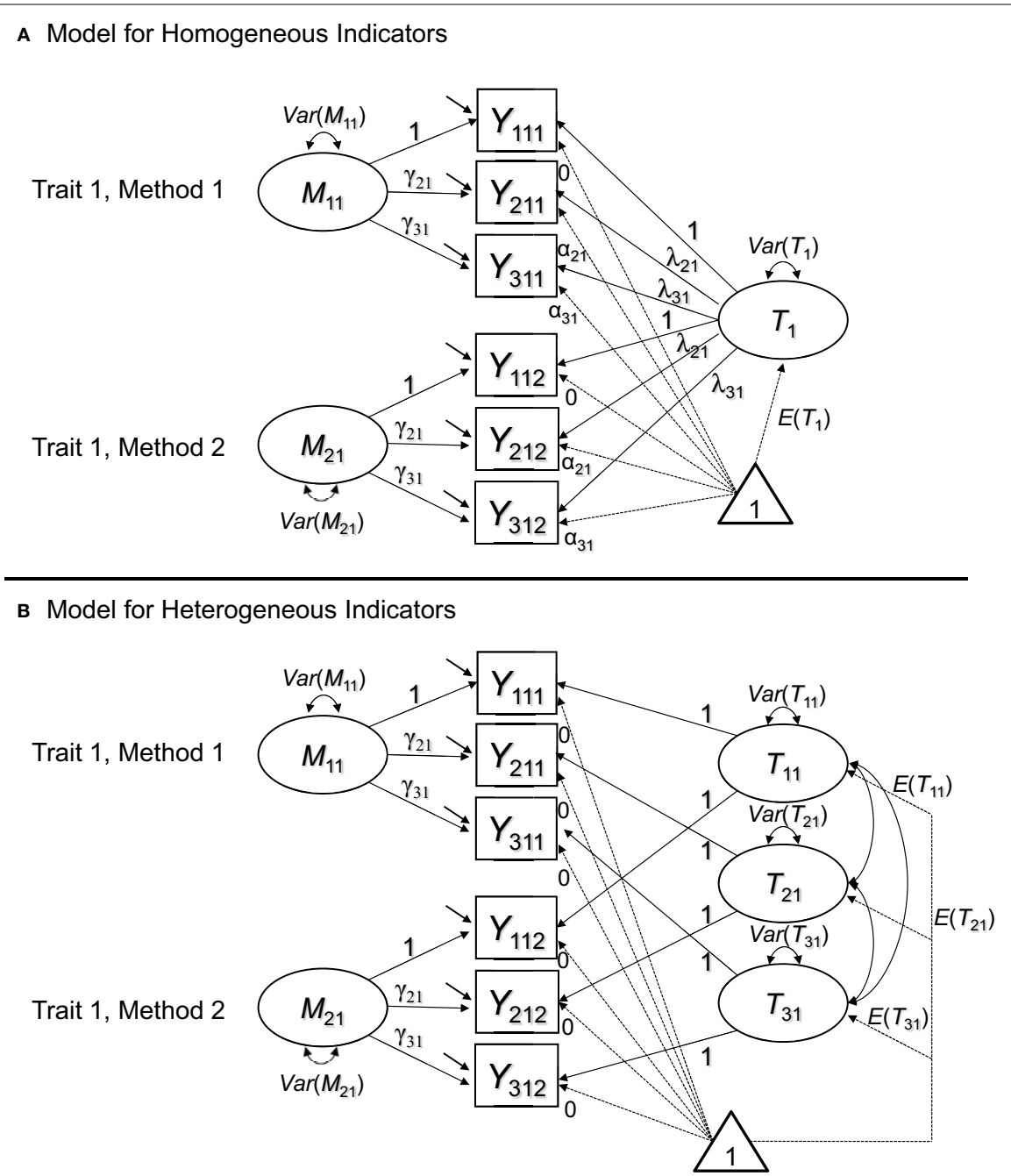


FIGURE 4 | CFA-MTMM models for interchangeable methods. (A) Version with a general trait factor for homogeneous indicators. **(B)** Version with indicator-specific traits for heterogeneous indicators. In the pictures, we show the recommended specification, in which loadings and intercepts are set equal across methods for identical indicators. In model version B, all trait factor loadings are fixed to 1 and all intercepts are fixed to zero.

person to come to a laboratory and bring two randomly selected friends to provide ratings of the person with respect to psychological variables (e.g., personality variables such as agreeableness etc.). The target person may not be sure whom to bring and may select his or her best friend as well as a more distant acquaintance. In this case, the "best friend" may have access to different information about the target person than the acquaintance. Hence, the two ratings may be considered structurally different rather than interchangeable. Non-interchangeable ratings may result in, for example, different latent means. This may result in a misfit of a model with equal loadings or intercepts, providing evidence against the assumed interchangeable nature of the two raters.

Strictly speaking, a researcher dealing with truly interchangeable methods (in the sense that the ratings represent random draws from a population of equivalent methods) would not only expect to find equal trait and method factor loadings as well as

equal intercepts across methods, but also equal error and equal method factor variances for each type of rater.

Obviously, the issue of MI in the CFA-MTMM model for interchangeable methods can often be resolved by making sure the methods are truly selected at random from a set of uniform methods. In this case, by definition, the measurement parameters are equivalent in the population (although they may differ in a sample due to sampling fluctuations). However, in practice, a random selection of raters or other methods may not always be feasible. Tests of MI can then provide a way to scrutinize whether the assumption of interchangeable methods is warranted or whether the chosen methods should better be treated as structurally different. In the latter case, the use of one of the three previously discussed models for structurally different methods would be preferable. Below we present an application of all five models to an MTMM study on ADHD symptoms.

EMPIRICAL ILLUSTRATION

SAMPLE AND MEASURES

The participants were mothers, fathers, and teachers of 1045 first grade children from 22 randomly selected elementary schools on the island of Majorca in the Balearic Islands and eight schools from Madrid (Spain). Assessments one and two occurred in the spring with assessment three occurring 12-months later. For the present illustration, we used data from the third assessment for which $N = 709$ (HI; $j = 1$) and $N = 710$ (IN; $j = 2$) cases with mother ratings ($k = 1$), father ratings ($k = 2$), and teacher ratings ($k = 3$) were available. The average age of the children was approximately 8 years with approximately 90% of the children being Caucasian and 10% North African.

Mothers, fathers, and teachers completed Child and Adolescent Disruptive Behavior Inventory (CADBI, Burns and Lee, 2010a,b). This study used the nine symptoms on the attention-deficit/hyperactivity disorder-inattention (ADHD-IN) and the nine symptoms on the ADHD-hyperactivity/impulsivity (HI) subscales. The ADHD symptoms were rated on a 6-point scale [i.e., *nearly occurs none of the time (e.g., 2 or fewer times per month)*, *seldom occurs (e.g., once per week)*, *sometimes occurs (e.g., a few times per week)*, *often occurs (e.g., once per day)*, *very often occurs (e.g., several times per day)*, and *nearly occurs all the time (e.g., many times per day)*].

For the purpose of this demonstration, item parcels were used as indicators rather than individual symptoms, given that earlier research provided justification for the use of parcels (Burns et al., in press).

MODELING STRATEGY

In Step 1 of our analyses, we attempted to establish a well-fitting baseline model for conducting subsequent MI analyses. For this purpose, we fit both Marsh and Hocevar's (1988) simple CFA model (**Figure 1**) and the extended model with indicator-specific factors (**Figure 2**) to the data and compared their fit to test whether homogeneity of the indicators (parcels) as well as configural invariance (equal factor structure) across raters could be assumed in the present application. None of the initial models included any formal equality constraints on measurement parameters. If the model in **Figure 1** for homogeneous indicators had fit the data well, it would have been the preferable model for further invariance tests relative to the more complex model in **Figure 2**, because the latter model is less parsimonious. In case of a substantially better fit of the more complex model in **Figure 2**, the more complex model is preferred, indicating a certain degree of indicator heterogeneity.

In Step 2, we proceeded with tests of MI across raters, using the best-fitting model from Step 1. The analyses in Step 2 began with a model of weak factorial invariance (only equal loadings across raters), then tested a strong invariance model (equal loadings and equal intercepts across raters), and finally a model of strict invariance (equal loadings, equal intercepts, and equal residual variances across raters). Given that all subsequent models were nested within previous models, we performed chi-square difference tests to compare the fit of the models directly. In cases in which one of the subsequent models showed a significantly worse fit than the preceding model, we further investigated issues of partial MI. That is, we tested in these cases whether there was invariance across some of the raters (e.g., mother and father, but not teacher ratings). Given that mother and fathers rated the children in the same context (at home), our hypothesis was that mother and father ratings may satisfy a stricter level of MI than parent and teacher ratings. In the final step, we tested for latent mean differences across raters if at least strong MI could be established for at least one pair of raters (e.g., mother and father ratings). In Step 3, we fit more complex CFA-MTMM models with method factors if this was warranted given the level of MI achieved in Steps 1 and 2.

All models were fit in Mplus 7 using maximum likelihood estimation. Examples of the Mplus specification for all models can be found in Appendix B in Supplementary Material. Global model fit was evaluated using the chi-square test, root mean square error of approximation (*RMSEA*), comparative fit index (*CFI*), Tucker-Lewis index (*TLI*), and standardized root mean square residual (*SRMR*). For a review and detailed discussion of these fit indices, see Schermelleh-Engel et al. (2003). Relative model fit was assessed via the chi-square difference test for nested models and Akaike's information criterion (*AIC*).

RESULTS OF THE MI ANALYSES

Table 1 shows global and relative model fit statistics for both the HI and IN constructs in the analyses of the **Figure 1**, **2** models. It can be seen that for both HI and IN, based on global model fit and the *AIC*, Model 1 (the configural invariance model without indicator-specific factors) clearly had to be rejected in favor of Model 2 (configural invariance with indicator-specific factors) as a baseline model. The configural invariance model with indicator-specific factors fit the data very well overall, showing a non-significant chi-square value for both HI and IN as well as excellent results based on other fit indices. An inspection of the model parameters revealed that for both HI and IN, the indicator-specific factors had significant (albeit relatively small) loadings (see **Table 2**), showing that the parcels were essentially, but not perfectly homogeneous. We therefore used the model with indicator-specific factors as the baseline model

Table 1 | Goodness of fit statistics for different models fit to the HI and IN multirater data set.

Model	χ^2	*df*	*p*	*RMSEA*	*CFI*	*TLI*	*SRMR*	$\chi^2\Delta$	*df* Δ	$p(\chi^2\Delta)$	*AIC*
HYPERACTIVITY/IMPULSIVITY											
Figure 1 configural invariance	269.91	24	<0.001	0.12	0.963	0.944	0.02				9773
Figure 2 configural invariance	17.70	17	0.41	0.01	1.000	1.000	0.01	—[a]	—[a]	—[a]	9535
Figure 2 weak invariance	21.40	21	0.44	0.01	1.000	1.000	0.02	3.7	4	0.45	9531
Figure 2 strong invariance	29.85	25	0.23	0.02	0.999	0.999	0.02	8.45	4	0.08	9531
Figure 2 strict invariance	52.60	31	0.009	0.03	0.997	0.996	0.02	22.75	6	<0.001	9542
Figure 2 strong invariance with equal means	104.27	27	<0.001	0.06	0.988	0.984	0.08	74.42[b]	2[b]	<0.001[b]	9602
Figure 2 strong invariance with equal means only for mother and father reports	**30.52**	**26**	**0.25**	**0.02**	**0.999**	**0.999**	**0.02**	**0.67[b]**	**1[b]**	**0.41[b]**	**9530**
INATTENTION											
Figure 1 configural invariance	196.64	24	<0.001	0.10	0.975	0.962	0.02				9386
Figure 2 configural invariance	16.09	17	0.52	0.00	1.000	1.000	0.01	—[a]	—[a]	—[a]	9219
Figure 2 weak invariance (all raters)	28.18	21	0.14	0.02	0.999	0.998	0.02	12.09	4	0.02	9223
Figure 2 weak invariance (mothers and fathers only)	17.66	19	0.55	0.00	1.000	1.000	0.01	1.57[c]	2[c]	0.46[c]	9217
Figure 2 strong invariance (all raters)	193.63	25	<0.001	0.10	0.976	0.965	0.04	165.45	4	<0.001	9381
Figure 2 strong invariance (mothers and fathers only)	29.20	23	0.17	0.02	0.999	0.999	0.02	11.54[d]	4[d]	0.02[d]	9220
Figure 2 strict invariance (all raters)	230.73	31	<0.001	0.10	0.971	0.966	0.04				9406
Figure 2 strict invariance (mothers and fathers only)	**30.39**	**26**	**0.25**	**0.02**	**0.999**	**0.999**	**0.02**	**1.19[e]**	**3[e]**	**0.76[e]**	**9215**
Figure 2 strict invariance across mothers and fathers only with equal means only for mother and father reports	34.81	27	0.14	0.02	0.999	0.998	0.02	4.42	1	0.04	9218

Note: RMSEA, root mean square error of approximation; CFI, comparative fit index; TLI, Tucker-Lewis index; SRMR, standardized root mean square residual; AIC, Akaike's information criterion. All chi-square difference tests refer to the previous model in the preceding row unless otherwise indicated. Bold-face indicates best-fitting models for which detailed results are presented.

[a] *No chi-square difference test reported, because model nesting involves boundary constraints in this case.*

[b] *Relative to the strong invariance model with unequal means.*

[c] *Relative to the configural invariance model.*

[d] *Relative to the model of weak invariance for mother and father reports only.*

[e] *Relative to the model of strong invariance for mother and father reports only.*

for subsequent MI tests involving equality constraints on loadings, intercepts, residual variances, and latent means for both HI and IN.

For HI, the assumptions of weak and strong MI across raters did not lead to a significant decline in model fit as indicated by chi-square difference tests[5]. The strong MI model also showed a very good global fit (non-significant chi-square). In contrast, the model of strict invariance was rejected by the chi-square difference test and also showed an increased *AIC* value relative to the strong invariance model. We concluded that error variances differed significantly across raters, whereas loadings and intercepts did not. Given that strong MI is sufficient for demonstrating scale equivalence and for meaningful comparisons of latent means, we proceeded with the strong-MI model and tested for latent mean differences across all three rater types for the HI construct.

The strong-MI model with equal means across all three rater types was clearly rejected for HI, showing that there were true mean differences across some of the raters (indicating a lack of convergent validity with respect to the true means or true mean differences between the home vs. school contexts). We additionally tested whether the means for mother and father ratings

[5]Indicator-specific loadings γ_{ijk} were not invariant across methods in the present application, indicating that different methods reflected indicator-specific effects differently. In particular, as can be seen from **Table 2**, teacher ratings showed much weaker (and partly insignificant) loadings on the IS_{ij} factors compared to parents, indicating that teachers did not differentiate as much between different facets of ADHD as did parents. This could potentially be explained by Halo effects that may have been more significant for teacher as compared to parent ratings in the present application.

Table 2 | Parameter estimates of the measurement models fit to the HI and IN multirater data set.

Parameter label	Hyperactivity/impulsivity ($j = 1$)			Inattention ($j = 2$)		
	Estimate	*SE*	Standardized estimate	Estimate	*SE*	Standardized estimate
TRAIT FACTOR LOADINGS						
λ_{1j}	1.00[a]	—	0.96[b]; 0.95[b]; 0.97[b]	1.00[a]	—	0.96[b]; 0.96[b]; 0.98[b]
λ_{2j}	0.93	0.01	0.90[b]; 0.90[b]; 0.94[b]	0.13	0.01	0.91[b]; 0.92[b]; 0.96[b]
λ_{3j}	0.92	0.01	0.90[b]; 0.90[b]; 0.94[b]	0.95	0.01	0.88[b]; 0.89[b]; 0.93[b]
INDICATOR-SPECIFIC FACTOR LOADINGS						
γ_{2j1}	1.00[a]	—	0.37	1.00[a]	—	0.33
γ_{2j2}	0.88	0.21	0.34	0.96	0.08	0.31
γ_{2j3}	0.17	0.06	0.07	0.31	0.07	0.09
γ_{3j1}	1.00[a]	—	0.24	1.00[a]	—	0.35
γ_{3j2}	1.57	0.56	0.40	0.94	0.08	0.32
γ_{3j3}	0.13	0.08	0.03	0.06	0.08	0.02
INTERCEPTS						
α_{1j1}	0.00[a]	—		0.00[a]	—	
α_{2j1}	−0.09	0.02		0.06	0.02	
α_{3j1}	0.08	0.02		0.16	0.02	
α_{1j2}	0.00[a]	—		0.00[a]	—	
α_{2j2}	−0.09	0.02		0.06	0.02	
α_{3j2}	0.08	0.02		0.16	0.02	
α_{1j3}	0.00[a]	—		0.00[a]	—	
α_{2j3}	−0.09	0.02		0.01	0.02	
α_{3j3}	0.08	0.02		−0.15	0.02	
ERROR VARIANCES						
$Var(\varepsilon_{1j1})$	0.09	0.01	0.08[c]	0.07	0.01	0.09[c]
$Var(\varepsilon_{2j1})$	0.07	0.04	0.06[c]	0.08	0.01	0.07[c]
$Var(\varepsilon_{3j1})$	0.15	0.02	0.13[c]	0.09	0.01	0.11[c]
$Var(\varepsilon_{1j2})$	0.10	0.01	0.09[c]	0.07	0.01	0.08[c]
$Var(\varepsilon_{2j2})$	0.09	0.03	0.08[c]	0.08	0.01	0.07[c]
$Var(\varepsilon_{3j2})$	0.02	0.06	0.02[c]	0.09	0.01	0.10[c]
$Var(\varepsilon_{1j3})$	0.07	0.01	0.06[c]	0.06	0.01	0.05[c]
$Var(\varepsilon_{2j3})$	0.11	0.01	0.11[c]	0.10	0.01	0.06[c]
$Var(\varepsilon_{3j3})$	0.13	0.01	0.12[c]	0.15	0.01	0.13[c]

Note: For hyperactivity/impulsivity, a model of strong invariance for all raters and equal means across mother and father ratings was chosen. For inattention, a model of strict invariance for mother and father ratings was chosen. λ_{ijk}*, trait factor loading (i, indicator; j, trait; k, method/rater);* γ_{ijk}*, indicator-specific factor loading;* α_{ijk}*, intercept;* $Var(\varepsilon_{ijk})$*, error variance. The methods used here are mother report (k = 1), father report (k = 2), and teacher report (k = 3).*

[a] *Parameter fixed for identification.*

[b] *Standardized loadings differed between raters for the same variable, because error variances and latent factor variances were allowed to differ in the final models. The standardized loadings are therefore given separately for each rater type in the following order: (1) mothers, (2) fathers, (3) teachers.*

[c] *Standardized residual variances indicate* $1 - R^2$ *and can be interpreted as coefficients of unreliability* $[(1 - Rel(Y_{ijk})]$ *for each variable.*

Dashes indicate fixed parameters for which no standard errors are computed.

were significantly different from one another, or whether the parent means only differed from the teacher means. A model with latent means constrained equal across mother and father (but not teacher) ratings was not rejected by the chi-square difference test relative to the strong MI model with unconstrained latent means. Therefore, we concluded that mothers and fathers did show convergent validity of mean levels for the HI construct, whereas the latent mean for teacher ratings was significantly smaller than for mothers and fathers. This indicated a lack of convergent validity with regard to the HI mean level across parent and teacher ratings or true differences in the mean HI levels between contexts (home vs. school; more details are provided later on when we discuss the parameter estimates of the final models).

Our analyses of the IN construct yielded different findings with regard to MI. In the IN case, already the weak invariance model showed a statistically significant (albeit relatively modest) increase in the chi-square relative to the configural invariance model, indicating at least partly non-invariant loadings across some of the raters. We tested whether the loadings were equivalent at least across mother and father ratings, as mothers and fathers were rating the same context (home). This hypothesis was not rejected by the chi-square difference test.

The strong-MI model for all raters showed a marked and highly significant increase in the chi-square relative to the full weak invariance model. We again tested whether strong invariance could at least be assumed across mothers and fathers. This hypothesis was also rejected according to the chi-square difference test, although the resulting chi-square difference was relatively modest and the global chi-square for this model was still non-significant. The strong-MI model for mothers and fathers also showed a very good global model fit, as indicated by a non-significant overall chi-square goodness of fit test. We therefore decided to proceed with the "partial strong-MI" model and to also test for strict MI across mother and father ratings only (leaving the intercept and residual variance parameters for teachers unconstrained). The strict-MI model for mothers and fathers showed a good overall model fit in terms of the chi-square and did not fit significantly worse than the partial strong-MI model. We therefore used the partial strict-MI model to test for mean differences across mothers and fathers. The resulting model with equal means across mothers and fathers showed a significant, albeit rather small chi-square difference value, indicating that there was a small difference in the latent means between mother and father ratings of IN.

A key finding of the IN analyses was the clear non-invariance of teacher intercepts relative to mother and father intercepts for this construct. The parameter estimates for the final models (strong MI across all three raters and equal latent means across mother and father reports for HI; strict MI across mother and father ratings only and unconstrained latent means across all raters) are presented in **Tables 2, 3**. **Table 2** contains the parameter estimates related to the measurement models (i.e., the loadings, intercepts, and error variances). **Table 3** shows the structural (latent variable model) parameter estimates (i.e., the latent covariances, correlations, variances, and means).

From **Table 2**, it can be seen that for IN, the intercepts for teacher ratings were markedly lower than the parent intercepts, indicating that teachers generally found it more "difficult" to endorse the symptoms on each of the IN indicators than did parents. One explanation could be that teachers in general are perhaps more used to seeing a broad spectrum of symptoms of IN and distraction in class than are parents at home. Therefore, the teachers in our sample may have used a different frame of reference (and a higher "threshold") when making their ratings of IN compared to parents. As a consequence, a more serious level of observed IN symptoms was required for teachers to produce the same score on the latent variable of IN as would be obtained from parent ratings. Interestingly, this difference was only found for IN, but not HI. This shows us that MI analyses in the context of MTMM data can reveal quite interesting information about differences between methods and how they may or may not use rating scale in a different way that may lead to scores that are not directly comparable. This information goes beyond what is typically assessed in MTMM studies and what can be obtained from an MTMM matrix alone.

From **Table 3**, we can see that there was substantial convergent validity in terms of the rank order of children for both HI and IN. Latent correlations ranged between 0.78 and 0.81 for mother and father ratings and between 0.42 and 0.45 between parents and teachers. Latent means for mother and father ratings of HI were set equal, given that this constraint was supported by the goodness-of-fit tests. In contrast, teacher ratings of HI resulted in a significantly smaller latent mean than parent ratings (0.71 vs. 1.10). The standardized mean difference was about 0.35, which can be seen as a small effect. It could be that teachers again use a different frame of reference for problems of HI, as they may be used to seeing a much larger array of problem behaviors at school than what most parents experience at home. Another explanation could be that the possibly more structured school context relative to less structure at home reduces the occurrence of HI symptoms in school relative to home for children within the normal range on HI.

For IN, the latent mean based on father reports was slightly larger than the mother-report mean (1.03 vs. 0.97, which represented a standardized mean difference below 0.10 and hence a very small effect). Our MI analyses for IN had indicated that the teacher mean could not be directly compared to the parent means, given the finding of intercept non-invariance for teacher as compared to parent ratings.

Table 3 | Estimated latent covariances, correlations, means, and variances in the final models.

	1.	2.	3.	4.	5.
HYPERACTIVITY/IMPULSIVITY					
1. T_{111}	—	0.85 (0.06)	0.45 (0.05)	—[a]	—[a]
2. T_{112}	0.81 (0.02)	—	0.43 (0.05)	—[a]	—[a]
3. T_{113}	0.42 (0.04)	0.42 (0.04)	—	—	—[a]
4. IS_{21}	—[a]	—[a]	—[a]	—	0.03 (0.01)
5. IS_{31}	—[a]	—[a]	—[a]	0.29 (0.07)	—
Means	1.10[b] (0.04)	1.10[b] (0.04)	0.71 (0.04)	—[a]	—[a]
Variances	1.11 (0.07)	0.99 (0.06)	1.05 (0.06)	0.16 (0.04)	0.07 (0.03)
INATTENTION					
6. T_{121}	—	0.60 (0.04)	0.41 (0.04)	—[a]	—[a]
7. T_{122}	0.78 (0.02)	—	0.44 (0.05)	—[a]	—[a]
8. T_{123}	0.45 (0.04)	0.44 (0.04)	—	—[a]	—[a]
9. IS_{22}	—[a]	—[a]	—[a]	—	0.05 (0.01)
10. IS_{32}	—[a]	—[a]	—[a]	0.44 (0.07)	—
Means	0.97 (0.04)	1.03 (0.04)	0.88 (0.04)	—[a]	—[a]
Variances	0.74 (0.05)	0.80 (0.05)	1.17 (0.07)	0.12 (0.02)	0.10 (0.02)

Note: Covariances are shown above the diagonal, correlations below the diagonal.

[a] *Covariances, correlations, or means that are set to zero by definition of the model. Standard errors are given in parentheses.*

[b] *Latent means for hyperactivity/inattention were set equal across mother and father reports.*

CT-C(M – 1) model

Our analyses with the CT-C($M-1$) model allowed us to estimate the consistency and method-specificity coefficients relative to a reference method. For the present example, we chose to select mother reports ($k = 1$) as reference method and contrast father

reports ($k = 2$) and teacher reports ($k = 3$) against this reference. This also made it possible to examine correlations between the father and teacher method factors, $Corr(M_{j2}, M_{j3})$. These correlations reflect whether fathers and teachers shared a common perspective that theses rater types did not share with mother reports.

Our results revealed high consistency and relatively low method-specificity coefficients for father reports for both HI and IN [range of consistency coefficients: $0.53 \leq Con(Y_{i12}) \leq 0.60$ for HI; $0.48 \leq Con(Y_{i22}) \leq 0.55$ for IN; range of method-specificity coefficients: $0.27 \leq MSpe(Y_{i12}) \leq 0.31$ for HI; $0.32 \leq MSpe(Y_{i22}) \leq 0.36$ for IN], indicating that there was high convergent validity between mother and father reports. This finding reflected the high correlations found between the mother and father trait factors in the baseline CFA model that we reported above. Consistency coefficients were lower (and method-specificity coefficients higher) for teacher ratings [range of consistency coefficients: $0.15 \leq Con(Y_{i13}) \leq 0.16$ for HI; $0.17 \leq Con(Y_{i23}) \leq 0.19$ for IN; range of method-specificity coefficients: $0.73 \leq MSpe(Y_{i13}) \leq 0.77$ for HI; $0.70 \leq MSpe(Y_{i23}) \leq 0.76$ for IN], showing that mother and teacher ratings shared less variance with each other than did mother and father ratings. The latent correlations between the father and teacher method factors were estimated to be $\varphi = 0.15$ for HI and $\varphi = 0.19$ for IN (both p-values were < 0.001). These correlations can be interpreted as partial correlations between father and teacher ratings after partialling out the common variance that both rater types shared with mother reports. In this case, the method factor correlations were rather small. This indicated that there was not much of a shared perspective between fathers and teachers above and beyond what fathers and teachers both shared with mothers.

Latent difference model

In the present example, a latent difference approach could be used for HI for all three rater types (mothers, fathers, and teachers), given that strong MI had been established across all three rater types for this construct. Given intercept non-invariance of teacher ratings as compared to parent ratings for IN, a latent difference approach would have been easily interpretable only for mother vs. father ratings for IN (excluding teacher ratings). We therefore only present the results for HI here as an example, for which we could meaningfully include all three rater types.

The latent difference model for HI yielded a latent difference factor mean of $E(T_{112} - T_{111}) = -0.02$ [$Var(T_{112} - T_{111}) = 0.40$] for father vs. mother ratings. This latent difference factor mean was not significantly different from zero ($p = 0.41$), showing that mother and father ratings of HI did not differ significantly in their latent means. The latent difference factor mean for teachers vs. mothers was estimated to be $E(T_{113} - T_{111}) = -0.40$ [$Var(T_{113} - T_{111}) = 1.26$], which was statistically significantly different from zero ($p < 0.001$). This again showed that mother and teacher ratings resulted in significantly different estimates of the overall level of HI in our data example. The correlations of father and teacher latent difference factors with the mother reference factor were $\varphi = -0.40$ and $\varphi = -0.56$, respectively. The latent difference factors for father and teacher ratings were correlated $\varphi = 0.33$.

Latent means model

The latent means model defines a common trait as the average across method-specific true score variables. Therefore, a latent means approach is only interpretable if *all* methods show at least strong MI. Otherwise, the "grand mean" of methods will be difficult to interpret. Therefore, we decided not to estimate the latent means model for IN because of intercept non-invariance for teacher ratings. The model was meaningful for HI, however, because all three rater types were shown to have invariant loadings and intercepts for this construct. The latent means model for HI yielded a common latent mean estimate of $E(T_1) = 0.97$ [$Var(T_1) = 0.73$]. The method factors in the latent means model represent deviations from the common latent mean factor. Their means indicate to which extent methods (on average) deviate from the grand mean across methods. The means of the method factors were estimated to be $E(M_{12}) = 0.12$ [$Var(M_{12}) = 0.21$] for father reports and $E(M_{13}) = -0.26$ [$Var(M_{13}) = 0.50$] for teacher reports. This reflected the fact that the latent mean of teacher ratings was substantially lower than the latent means for mother and father ratings of HI. The common trait was correlated $\varphi = 0.06$ with M_{12} and $\varphi = -0.15$ with M_{13}. The correlation between M_{12} and M_{13} was estimated to be $\varphi = -0.73$.

CFA-MTMM model for interchangeable methods

We fit both the general and the indicator-specific trait versions of Eid et al.'s (2008) CFA-MTMM model for interchangeable methods to our data example to test whether the ratings would satisfy the restrictions implied by the interchangeable model (i.e., invariant loadings and intercepts as shown in **Figures 4A,B**). Note that from a measurement theoretical point of view, mother, father, and teacher ratings would typically not be seen as interchangeable methods, because they are not sampled from the same "universe" of methods; here, we use these data merely to illustrate MI analyses in the interchangeable MTMM model and do not imply that the ratings should be treated as interchangeable.

The fit statistics are presented in **Table 4**. Parameter estimates for the final models are given in **Table 5**. We found that a model with invariant loadings and intercepts for all three types of raters (mother, father, and teacher) was not tenable for either HI or IN, even if the less restrictive version of the model with indicator-specific traits was used. One reason was that in this model, the latent means are implicitly assumed to be equal across all interchangeable methods—this assumption was already rejected in our initial analyses of the simple CFA model.

In contrast, an indicator-specific trait model for mother and father reports only (dropping teacher reports from the analysis) fit both the HI and IN data well, showing that mother and father ratings satisfied the conditions of interchangeability implied by the model in this application. This parallels our findings from the previous analyses according to which mother and father ratings were more similar to one another than they were compared to teacher ratings. For both HI and IN, mothers and fathers shared the same or very similar means as indicated by previous analyses.

Table 4 | Goodness of fit statistics for different versions of the CFA-MTMM model for interchangeable methods fit to the HI and IN multirater data set.

Model	χ^2	*df*	*p*	*RMSEA*	*CFI*	*TLI*	*SRMR*	$\chi^2\Delta$	*df*Δ	$p(\chi^2\Delta)$	*AIC*
HYPERACTIVITY/IMPULSIVITY											
All three raters; equal loadings and intercepts	378.01	31	<0.001	0.13	0.948	0.939	0.16				9867
Mothers and fathers only; equal loadings and intercepts	7.34	8	0.50	0.00	1.000	1.000	0.01				6102
Mothers and fathers only; equal loadings, intercepts, and residual variances	**12.70**	**11**	**0.31**	**0.02**	**1.000**	**0.999**	**0.01**	**5.36**	**3**	**0.15**	**6102**
Mothers and fathers only; equal loadings, intercepts, residual variances, and method factor variances	16.81	12	0.16	0.03	0.999	0.999	0.03	4.11	1	0.04	6104
INATTENTION											
All three raters; equal loadings and intercepts	381.16	31	<0.001	0.13	0.949	0.941	0.09				9556
Mothers and fathers only; equal loadings and intercepts	9.42	8	0.31	0.02	1.000	0.999	0.01				5648
Mothers and fathers only; equal loadings, intercepts, and residual variances	10.10	11	0.52	0.00	1.00	1.00	0.01	0.68	3	0.88	5643
Mothers and fathers only; equal loadings, intercepts, residual variances, and method factor variances	**11.15**	**12**	**0.52**	**0.00**	**1.000**	**1.000**	**0.02**	**1.05**	**1**	**0.31**	**5642**

*Note: In order to save space, we only present results for the model version with indicator-specific traits (**Figure 4B**), given that the model version with a single trait (**Figure 4A**) did not fit well for any rater combination. RMSEA, root mean square error of approximation; CFI, comparative fit index; TLI, Tucker-Lewis index; SRMR, standardized root mean square residual; AIC, Akaike's information criterion. For both constructs, the initial model included all three rater types. Subsequent models included only mother and father ratings (dropping teacher ratings from the analysis). Bold-face indicates best-fitting models.*

(The baseline model for IN yielded a significant mean difference between mother and father reports; this was likely because there was more statistical power to detect mean differences in the combined model with all three raters. Nonetheless, the mean difference between mothers and fathers was very small also in the initial analysis).

We also tested more strict models of interchangeability for mother and father ratings, in which we also constrained (a) the error variances and (b) the method factor variances to be equal across mother and father ratings. Chi-square difference tests revealed that for HI, equal error variances were tenable, but not equal method factor variances. In contrast, for IN, both the assumption of equal error variances and the assumption of equal method factor variances were acceptable according to the chi-square difference test. In summary, mother and father ratings of IN could be viewed as strictly interchangeable in the sense of the model, whereas for HI the ratings could be viewed as interchangeable except for the amount of method factor variance.

DISCUSSION

Researchers frequently use different raters as methods in MTMM studies. Often, ratings are provided on comparable items or scales. In these cases, one can examine whether (1) different raters use the items or scales in equivalent ways (i.e., whether MI holds across methods) and (2) whether there is convergent validity of latent mean levels across methods. This opens up new possibilities for studying method effects in more detail. Whereas traditional approaches to MTMM analyses (such as Campbell and Fiske's, 1959; MTMM matrix or conventional CFA-MTMM models) have typically focused exclusively on (observed or latent) relationships (correlations) between *different* TMUs, the MI approach presented here first of all examines the relationships between indicators and latent variables *within* each TMU. In this article, we proposed to analyze MI using a baseline MTMM model without method factors in the first step of the analysis. Using this model, researchers can first of all test whether the proposed factor structure holds within corresponding TMUs and second, whether the way indicators relate to latent factors within a TMU is comparable across methods. This allows researchers to examine whether supposedly equivalent concepts that are measured by the same indicators (but different methods) have similar relationships with their indicators for different methods. (This may be termed an examination of the "convergent validity of measurement properties."). If they do, this may increase a researcher's confidence that similar concepts are indeed measured by each method or at least that the indicators "function" similarly across methods. If they don't, then a researcher may question whether the concepts can be seen as equivalent across methods, warranting further study

Table 5 | Parameter estimates in the CFA-MTMM models for interchangeable methods fit to mother and father ratings of HI and IN.

Parameter label	Hyperactivity/impulsivity ($j = 1$)			Inattention ($j = 2$)		
	Estimate	*SE*	Standardized estimate	Estimate	*SE*	Standardized estimate
TRAIT FACTOR LOADINGS						
λ_{1j}	1.00[a]	—	0.84; 0.89[b]	1.00[a]	—	0.84
λ_{2j}	1.00[a]	—	0.88; 0.91[b]	1.00[a]	—	0.87
λ_{3j}	1.00[a]	—	0.84; 0.88[b]	1.00[a]	—	0.85
METHOD FACTOR LOADINGS						
γ_{1j}	1.00[a]	—	0.37	1.00[a]	—	0.47
γ_{2j}	0.88	0.05	0.34	1.05	0.05	0.41
γ_{3j}	0.99	0.05	0.07	0.98	0.05	0.43
ERROR VARIANCES						
$Var(\varepsilon_{1j})$	0.09	0.01	0.08; 0.08[b,c]	0.07	0.01	0.08[c]
$Var(\varepsilon_{2j})$	0.09	0.01	0.07; 0.08[b,c]	0.09	0.01	0.08[c]
$Var(\varepsilon_{3j})$	0.09	0.01	0.08; 0.09[b,c]	0.09	0.01	0.10[c]
LATENT MEANS						
$E(T_{1j})$	1.09	0.04		0.98	0.03	
$E(T_{2j})$	0.95	0.04		1.17	0.04	
$E(T_{3j})$	1.08	0.04		1.09	0.04	
LATENT VARIANCES						
$Var(T_{1j})$	0.86	0.06		0.58	0.04	
$Var(T_{2j})$	0.92	0.06		0.88	0.06	
$Var(T_{3j})$	0.78	0.06		0.66	0.05	
$Var(M_{j1})$	0.26	0.04		0.18[d]	0.02	
$Var(M_{j2})$	0.14	0.03		0.18[d]	0.02	

For hyperactivity/impulsivity, a model of strong invariance for all raters and equal means across mother and father ratings was chosen. For inattention, a model of strict invariance for mother and father ratings was chosen. λ_{ijk}, trait factor loading (i, indicator; j, trait; k, method/rater); γ_{ijk}, indicator-specific factor loading; α_{ijk}, intercept; $Var(\varepsilon_{ijk})$, error variance. The methods used here are mother report (k = 1), father report (k = 2), and teacher report (k = 3).

[a] Parameter fixed for identification.

[b] Standardized loadings and standardized error variances for HI differed between raters for the same variable, because the method factor variances were allowed to differ in the final model. The standardized loadings and error variances are therefore given separately for each rater type in the following order: (1) mothers, (2) fathers.

[c] Standardized residual variances indicate $1 - R^2$ and can be interpreted as coefficients of unreliability [$(1 - Rel(Y_{ijk}))$] for each variable.

[d] Method factor variances were set equal across mother and father reports in this model.

Dashes indicate fixed parameters for which no standard errors are computed.

of what exactly is measured by each method and in which ways concepts or item meanings might differ across methods.

Non-invariant intercepts or loadings across methods can indicate that the scales have a different meaning for different methods. For example, a certain behavior may be less relevant for the definition of a construct for one type of rater than for another. Consider, for instance, symptoms of ADHD. A specific ADHD-IN symptom may be highly relevant for parents' view of their children, but not so critical for teachers' view (maybe because it does not occur in the school context), thus leading to different factor loadings or intercepts for the same symptom. The finding of measurement non-invariance can thus shed more light on how different indicators (e.g., ADHD symptoms) "function" for different types of raters.

It is interesting to note that our approach of beginning an MTMM analysis with a thorough investigation of the measurement properties within and across TMUs seems to be well in line with Fiske and Campbell's (1992) later view of the original Campbell and Fiske (1959) MTMM approach. In their 1992 review, Fiske and Campbell proposed to "settle for the practice of studying 'TMUs'," given that "method and trait or content are highly interactive and interdependent" (p. 394). Examining MI across TMUs is one component of such an analysis strategy that places more emphasis on what is measured within each TMU rather than just on correlations between TMUs.

While traditional MTMM analyses focus exclusively on correlation or covariance structures (Campbell and Fiske, 1959), we propose to routinely consider means in the analyses as well, which is a novel aspect in MTMM research. By including means in the analysis, MI across methods can be more fully tested and, if strong MI can be established, latent means can be compared across methods to examine the degree of convergent validity of mean levels across methods. When using just the covariance matrix (and no means), researchers can test for loading (weak factor) invariance and invariant error variances, but not for intercept (strong) MI. In our empirical example, we found that the

intercepts were non-invariant across some methods for one of the constructs, indicating differences in scale difficulty across methods. This information could not have been obtained without including means in the analysis.

In addition, when only covariances or correlations are analyzed, latent means cannot be compared across methods. Strong MI is a prerequisite for interpreting latent mean differences across methods in a straightforward way. With non-invariant intercepts and/or loadings, latent mean differences across methods may be difficult to interpret, because the measurements would not be in the same metric in this case. In cases of non-invariance, mean differences would represent a mixture between rater biases and measurement bias (differences in scale use). This was the case for parent vs. teacher ratings of IN in our empirical example. Of course, the question of interpretability depends on the particular substantive application and is also a matter of degree rather than "all or nothing." For example, if violations of MI are small, approximate MI (van de Schoot et al., 2013) may still hold, warranting a proper interpretation of latent mean differences across methods.

If strong MI can be established across methods (such as in our empirical application to the HI construct), then a researcher can meaningfully test whether methods converge in the assessment of the latent mean level of a construct in a given population. In our HI example, this was the case for mothers vs. fathers, but not for parents vs. teachers. This showed that there was a lack of convergent validity of mean levels across methods (or a true difference between the home and school contexts), even though the convergent validity in terms of the correlations between parents and teachers was quite strong. This issue is especially critical when researchers want to draw conclusions about, for example, the overall level of a problem such as HI. In this case, they would come to different conclusions based on parent vs. teacher ratings in the present example. It is therefore important to examine the convergent validity of mean levels across raters in such cases.

Of course, testing for MI and comparing latent means only makes sense when methods used comparable measurement instruments (items and response scales) to begin with. When very different methods are used (e.g., ratings vs. physiological measures of stress), tests of MI are typically not meaningful, especially when scoring procedures differ between methods (e.g., 4-point Likert scale vs. cortisol concentration in nmol/L). When different methods used similar response scales as in the examples presented in this paper, but strong MI cannot be established, observed mean differences for corresponding indicators across methods still provide valuable information, as they indicate method effects at the measurement level (i.e., differences in scale use; see discussion above).

DIFFERENT MODELS

In this article, we demonstrated that including mean structures and testing for MI is not only an issue of potential substantive interest in MTMM analyses. With respect to more complex CFA-MTMM models with method factors, we showed that MI is relevant to these models especially for two reasons: (1) the definition of trait and method factors may require strong MI for a meaningful interpretation of structural parameters such as latent trait and method factor means and variances as well as individual scores on these latent variables or (2) at least strong MI is implied by a CFA model for interchangeable methods. Therefore, MI is not just something that researchers may or may not be interested in when analyzing MTMM data; instead, depending on the model, MI can be a prerequisite for the proper interpretation of one's MTMM model or for conclusions about whether methods can be seen as interchangeable or not.

We showed that among the more complex models for structurally different methods discussed here, the CT-C($M-1$) model makes the least restrictive assumptions in terms of MI. That is, for calculating coefficients of convergent validity (consistency) and method specificity in this model or for the interpretation of trait and method factors in general, MI beyond configural invariance is not required. The only case in which MI can become relevant in the CT-C($M-1$) model is when researchers want to interpret unstandardized structural regression coefficients or latent mean differences derived from these coefficients.

In contrast, the latent difference and latent means models require MI across methods by definition. When different methods used comparable items that were measured on the same scale (or rescaled to the same metric) and provided that strong MI across methods can be established, then the latent difference and latent means models provide a meaningful and straightforward estimation of mean method effects, because means can be directly estimated for the method factors in these models. In contrast, in the CT-C($M-1$) model, mean method effects are not directly estimated in terms of method factor means, because the method factors in this model have means of zero by definition. Nonetheless, mean method effects can also be analyzed in the CT-C($M-1$) model as shown in detail in Geiser et al. (2012). An advantage of the CT-C($M-1$) model is that it can be used even when different methods used completely different metrics (e.g., self-reports on a 4-point Likert scale vs. cortisol concentrations as measures of stress) or when MI does not hold.

The latent difference model is less restrictive with regard to MI than the latent means model, as the latent difference model can still be used in cases of partial MI (when at least one non-reference method shows strong MI relative to the reference method). In contrast, a proper interpretation of the common trait in the latent means model requires that strong MI between all methods be established. Despite these differences between the CT-C($M-1$), latent difference, and latent means models, all three models provide meaningful definitions of trait and method factors. The choice of a particular model depends in part on a researcher's specific goals in a given application.

Another area of MTMM research for which MI plays a role is when researchers study interchangeable methods. We considered Eid et al.'s (2008) CFA-MTMM model for interchangeable methods separately, because it is designed for a different data structure (interchangeable methods) than the CT-C($M-1$), latent difference, and latent means models (which are designed for structurally different methods). If, for example, a researcher wants to test whether methods that, based on theory, are conceived of as interchangeable truly are interchangeable in a statistical sense, he or she should use an appropriate CFA-MTMM model for interchangeable methods and test for MI. If this assumption is rejected,

the methods in question may be better viewed as structurally different. In this case, one of the models for structurally different methods [CT-C($M-1$), latent difference, or latent means] may be more appropriate.

CONCLUSION

Most MTMM studies so far have focused on assessing convergent validity in terms of correlations between methods or raters selected to measure the same constructs. We argued that useful incremental information about method effects can be gained from including mean structures in MTMM models and testing for MI across methods. Furthermore, we showed that MI is relevant in more complex CFA-MTMM models with method factors, either because the definitions of trait and method factors imply MI for a meaningful interpretation of structural parameters or because the type of method (interchangeable vs. structurally different) may or may not imply MI across methods. We hope that researchers will find our article useful as a guide for future, more fine-grained studies of method effects.

ACKNOWLEDGMENTS

This research was funded by a grant from the National Institutes on Drug Abuse (NIH-NIDA), grant #1 R01 DA034770-01 awarded to Ginger Lockhart and Christian Geiser. The data used in the empirical application was collected under research grant #PSI2011-23254 from the Spanish government awarded to Mateu Servera and G. Leonard Burns. We would like to thank the editor, the two reviewers who provided open reviews, and a third reviewer who decided to remain anonymous for providing valuable feedback and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.01216/abstract

REFERENCES

Burns, G. L., and Lee, S.-Y. (2010a). *Child and Adolescent Disruptive Behavior Inventory—Parent Version*. Pullman, WA: Author.

Burns, G. L., and Lee, S.-Y. (2010b). *Child and Adolescent Disruptive Behavior Inventory—Teacher Version*. Pullman, WA: Author.

Burns, G. L., Servera, M., del Mar Bernard, M., Carrillo, J. M., and Geiser, C. (in press). Mothers', fathers', teachers', and aides' ratings of ADHD symptoms and academic impairment: implications for DSM-5 ADHD diagnostic criterion C. *Psychol. Assess.*

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

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Cole, D. A., and Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: a confirmatory factor analytic approach. *Multivariate Behav. Res.* 20, 389–417. doi: 10.1207/s15327906mbr2004_3

Cole, D. A., Truglio, R., and Peeke, L. (1997). Relation between symptoms of anxiety and depression in children: a multitrait-multimethod-multigroup assessment. *J. Consult. Clin. Psychol.* 65, 110–119. doi: 10.1037//0022-006X.65.1.110

Dumenci, L. (2000). "Multitrait-multimethod analysis," in *Handbook of Applied Multivariate Statistics and Mathematical Modeling,* eds S. D. Brown and H. E. A. Tinsley (San Diego, CA: Academic Press), 583–611. doi: 10.1016/B978-012691360-6/50021-5

Eid, M., Lischetzke, T., and Nussbeck, F. W. (2006). "Structural equation models for multitrait-multimethod data," in *Handbook of Multimethod Measurement in Psychology*, eds M. Eid and E. Diener (Washington, DC: American Psychological Association), 283–299. doi: 10.1037/11383-020

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika* 65, 241–261. doi: 10.1007/BF02294377

Eid, M., Lischetzke, T., Nussbeck, F. W., and Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: a multiple indicator CT-C(M–1) model. *Psychol. Methods* 8, 38–60. doi: 10.1037/1082-989X.8.1.38

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., and Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: different models for different types of methods. *Psychol. Methods* 13, 230–253. doi: 10.1037/a0013219

Eid, M., Schneider, C., and Schwenkmezger, P. (1999). Do you feel better or worse? The validity of perceived deviations of mood states from mood traits. *Eur. J. Pers.* 13, 283–306.

Fiske, D. W., and Campbell, D. T. (1992). Citations do not solve problems. *Psychol. Bull.* 112, 393–395. doi: 10.1037/0033-2909.112.3.393

Geiser, C., Bishop, J., Lockhart, G., Shiffman, S., and Grenard, J. (2013). Analyzing latent state-trait and multiple-indicator latent growth curve models as multi-level structural equation models. *Front. Psychol.* 4:975. doi: 10.3389/fpsyg.2013.00975

Geiser, C., Eid, M., and Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C(M–1) model: a comment on Maydeu-Olivares and Coffman (2006). *Psychol. Methods* 13, 49–57. doi: 10.1037/1082-989X.13.1.49

Geiser, C., Eid, M., West, S. G., Lischetzke, T., and Nussbeck, F. W. (2012). A comparison of method effects in two confirmatory factor models for structurally different methods. *Struct. Equ. Model.* 19, 409–436. doi: 10.1080/10705511.2012.687658

Geiser, C., Koch, T., and Eid, M. (2014a). Data-generating mechanisms versus constructively-defined latent variables in multitrait-multimethod analyses: a comment on Castro-Schilo, Widaman, & Grimm (2013). *Struct. Equ. Model.* doi: 10.1080/10705511.2014.919816. (in press)

Geiser, C., Mandelman, S. D., Tan, M., and Grigorenko, E. L. (2014b). Multitrait-multimethod assessment of giftedness: an application of the correlated traits-correlated (Methods – 1) model. *Struct. Equ. Model.* (in press)

Grigorenko, E. L., Geiser, C., Slobodskaya, H. R., and Francis, D. J. (2010). Cross-informant symptoms from CBCL, TRF, and YSR: trait and method variance in a normative sample of Russian youths. *Psychol. Assess.* 22, 893–911. doi: 10.1037/a0020703

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: practical and theoretical issues. *Multivariate Behav. Res.* 32, 53–76. doi: 10.1207/s15327906mbr3201_3

Marsh, H. W., Byrne, B. M., and Craven, R. (1992). Overcoming problems in confirmatory factor analyses of MTMM data: the correlated-uniqueness model and factorial invariance. *Multivariate Behav. Res.* 27, 489–507. doi: 10.1207/s15327906mbr2704_1

Marsh, H. W., and Hocevar, D. (1988). A new, more powerful approach to multitrait-multimethod analyses: application of second-order confirmatory factor analysis. *J. Appl. Psychol* 73, 107–117. doi: 10.1037/0021-9010.73.1.107

McArdle, J. J. (1980). Causal modeling applied to psychonomic systems simulation. *Behav. Res. Methods* 12, 193–209. doi: 10.3758/BF03201598

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (1995). "The statistical analysis of method effects in multitrait-multimethod data: a review," in *Personality Research, Methods, and Theory: A Festschrift Honoring Donald W. Fiske,* eds P. E. Shrout and S. T. Fiske (Hillsdale, NJ: Erlbaum), 93–109.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Nussbeck, F. W., Eid, M., Geiser, C., Courvoisier, D. S., and Lischetzke, T. (2009). A CTC(M-1) model for different types of raters. *Methodology* 5, 88–98. doi: 10.1027/1614-2241.5.3.88

Pohl, S., and Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behav. Res.* 45, 45–72. doi: 10.1080/00273170903504729

Pohl, S., Steyer, R., and Kraus, K. (2008). Modelling method effects as individual causal effects. *J. R. Stat. Soc. A* 171, 41–63.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behav. Res.* 47, 667–696. doi: 10.1080/00273171.2012.715555

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003). Evaluating the fit of structural equation models: test of significance and descriptive goodness-of-fit measures. *Methods Psychol. Res*. 8, 23–74.

Servera, M., Lorenzo-Seva, U., Cardo, E., Rodríguez-Fornells, A., and Burns, G. L. (2010). Understanding trait and source effects in ADHD and ODD rating scales: Mothers', fathers' and teachers' ratings of children from the Balearic Islands. *J. Clin. Child Adolesc. Psychol*. 39, 1–11. doi: 10.1080/15374410903401187

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthen, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

Vautier, S., Raufaste, E., and Cariou, M. (2003). Dimensionality of the Revised Life Orientation test and the status of filler items. *Int. J. Psychol*. 38, 390–400. doi: 10.1080/00207590344000222

Widaman, K.-F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Appl. Psychol. Meas.* 9, 1–26. doi: 10.1177/014662168500900101

Widaman, K. F., and Reise, S. P. (1997). "Exploring the measurement invariance of psychological instruments: applications in the substance use domain," in *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research,* eds K. J. Bryant, M. Windle, and S. G. West (Washington, DC: American Psychological Association), 281–324.

Woehr, D. J., Sheehan, M. K., and Bennett, W. (2005). Assessing measurement equivalence across rating sources: a multitrait-multirater approach. *J. Appl. Psychol*. 90, 592–600. doi: 10.1037/0021-9010.90.3.592

Wu, A. D., Li, Z., and Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: a demonstration with TIMSS data. *Pract. Assess. Res. Eval.* 12. Available online at: http://pareonline.net/getvn.asp?v=12&n=3

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 February 2014; accepted: 07 October 2014; published online: 30 October 2014.

*Citation: Geiser C, Burns GL and Servera M (2014) Testing for measurement invariance and latent mean differences across methods: interesting incremental information from multitrait-multimethod studies. Front. Psychol. **5**:1216. doi: 10.3389/fpsyg.2014.01216*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in
PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 17 September 2014
doi: 10.3389/fpsyg.2014.00980

The consequences of ignoring measurement invariance for path coefficients in structural equation models

Nigel Guenole [1,2] **** and Anna Brown*** [3]

[1] *IBM Smarter Workforce Institute, London, UK*
[2] *Institute of Management Studies, Goldsmiths, University of London, London, UK*
[3] *School of Psychology, Keynes College, University of Kent, Canterbury, UK*

Edited by:
Rens Van De Schoot, Utrecht University, Netherlands

Reviewed by:
Jelte M. Wicherts, Tilburg University, Netherlands
Peter Lugtig, Utrecht University, Netherlands

****Correspondence:***
Nigel Guenole, Institute of Management Studies, Goldsmiths, University of London, New Cross, London, SE14 6NW, UK
e-mail: n.guenole@gold.ac.uk

We report a Monte Carlo study examining the effects of two strategies for handling measurement non-invariance – modeling and ignoring non-invariant items – on structural regression coefficients between latent variables measured with item response theory models for categorical indicators. These strategies were examined across four levels and three types of non-invariance – non-invariant loadings, non-invariant thresholds, and combined non-invariance on loadings and thresholds – in simple, partial, mediated and moderated regression models where the non-invariant latent variable occupied predictor, mediator, and criterion positions in the structural regression models. When non-invariance is ignored in the latent predictor, the focal group regression parameters are biased in the opposite direction to the difference in loadings and thresholds relative to the referent group (i.e., lower loadings and thresholds for the focal group lead to overestimated regression parameters). With criterion non-invariance, the focal group regression parameters are biased in the same direction as the difference in loadings and thresholds relative to the referent group. While unacceptable levels of parameter bias were confined to the focal group, bias occurred at considerably lower levels of ignored non-invariance than was previously recognized in referent and focal groups.

Keywords: measurement equivalence/invariance, categorical indicators, structural equation modeling

INTRODUCTION

Methodologists in the social sciences differentiate between two primary forms of psychometric equivalence, or ways of showing that a psychometric instrument functions similarly in a probabilistic sense across multiple populations. Measurement invariance exists when two individuals sampled from different sub-populations but with the same standing on the latent continuum have the same expected test score (Drasgow, 1982, 1984; Mellenbergh, 1989; Meredith, 1993; Vandenberg and Lance, 2000; Kankaraš et al., 2011; Davidov et al., 2014). Today measurement invariance is considered a fundamental issue in psychological testing (Lubke et al., 2003) that has social as well as statistical consequences (Beuckelaer et al., 2007; Borsboom et al., 2008). In studies of measurement invariance, the groups under study are designated as either the referent or focal group (Holland and Thayer, 1988). Next, the equivalence of measurement model parameters, usually the item loadings and intercepts or thresholds, is examined using approaches based on either item response theory (IRT) or multiple group confirmatory factor analysis (CFA). In the CFA approach, which is the focus of the present article, a series of competing models is fitted to response data, where the group membership acts as a potential categorical moderator (e.g., French and Finch, 2011). Equivalent measurement model parameters across groups are required for comparable measurement, a consideration identical to the use of equal measurement scales (say, degrees centigrade) when comparing temperatures in two different regions. For a recent description of the process of examining measurement invariance see French and Finch (2011) or van de Schoot et al. (2012).

Relational invariance, the second form of equivalence, examines whether the same structural relationships hold between variables across two or more subpopulations (Mellenbergh, 1989; Meredith, 1993). When variables under study are latent, the slopes of structural regression paths in multiple group analyses are examined for invariance[1]. Drasgow (1984) has argued that there is a logical sequence to testing measurement equivalence: measurement invariance should first be tested, followed by relational invariance. If non-invariance is observed in the measurement model, the researcher might want to delete "offending" items. This might not be appropriate if the questionnaire is a well-established instrument. Remaining options include freely estimating the parameters for the non-invariant items to achieve partial invariance (Byrne et al., 1989), or to ignore the non-invariance. The challenge faced by the researcher who allows partial invariance is how much non-invariance can be tolerated whilst still claiming that the same construct is measured across groups or between current and past research. The challenge faced by the researcher ignoring the non- invariance is whether the results of the misspecified model can be trusted. In practice, applied researchers should make a decision based on the expected threats to the validity of their conclusions under each course of action.

Sometimes, the primary focus of the researcher is to examine structural relations across groups of interest. Wasti et al.

[1] There is a well-developed literature on relational equivalence where regressions are conditioned on the observed variable composite. See, for example Drasgow (1982, 1984) and the duality theorems of Millsap (1995, 1998).

(2000) for example, examined whether the antecedents and consequences of sexual harassment were the same between the United States and Turkey. If it was certain that ignoring measurement non-invariance across populations would lead to negligible differences in relationships between latent variables, it could be tempting to do so. On the other hand, it would be necessary to model the non-invariance if ignoring it would result in a substantial regression parameter bias. There have been at least three calls for Monte Carlo studies of such issues in the literature (Chen, 2008; Schmitt and Kuljanin, 2008; Schmitt et al., 2011). The present article addresses this call for a Monte Carlo study of measures employing categorical indicators. Our approach broadly follows the recommendations of Paxton et al. (2001) and Boomsma (2013).

PAST RESEARCH ON THE EFFECT OF MEASUREMENT INVARIANCE ON RELATIONAL INVARIANCE

Chen (2008) reported a Monte Carlo investigation of the impact of ignoring measurement non-invariance in the slopes of linear factor models on the relative bias in regression parameters of structural models. She found that when referent group loadings were higher on the exogenous latent variable, the referent group regression parameter was overestimated, i.e., the relative bias was positive, and the regression parameter in the focal group was underestimated, i.e., the relative bias was negative. The pattern was reversed when the non-invariant construct was the latent criterion variable. The relative bias in the regression parameters was always greater in the focal group. However, extreme levels of non-invariance had to be ignored before adverse effects on regression coefficient accuracy emerged.

Oberski (2014) used Monte Carlo studies to examine the expected change in the parameter of interest statistic (EPC-Interest: Satorra, 1989; Bentler and Chou, 1992) as a method for examining the sensitivity of parameters under study to misspecification of invariance constraints. This method has the advantage of avoiding the unnecessary rejection of the measurement invariance model, and alerting the researcher to doubtful substantive conclusions about parameters when measurement invariance appears to hold. Unlike the more familiar expected parameter change (EPC: Saris et al., 1987), EPC-Interest examines the change in parameters of interest other than the parameter being fixed or freed. Obserski examined changes in regression parameters of a random effects model due to ignoring versus modeling non-invariant loadings. The effects on the regression coefficient in the empirical example used in that article were generally small.

THEORETICALLY DERIVED RESEARCH QUESTION

We extend the work of Chen (2008) and Oberski (2014) in several new directions. While Oberski (2014) evaluated a method for examining the impact of the non-invariance problem in specific models, this study examines the extent of these effects in general structural relationships under typical conditions. Whereas Chen examined the impact of measurement non-invariance on simple regression parameters, structural models in practice are usually more complex. We examined the effect of ignoring non-invariance on partial regression coefficients, i.e., regression with covariates, mediated regression coefficients, and moderated regression coefficients. In each case, we examined the effect of ignoring the invariance when the latent variable with non-invariant parameters was the predictor, when it was the criterion, and when the latent variable with non-invariance occupied the mediator position in the model.

The current investigation extends the work of Chen (2008) and Oberski (2014) in a further important way. We generalize these authors' earlier results to models incorporating categorical factor indicators, thus focusing on loadings and thresholds in categorical item factor analyses (CIFA: Forero and Maydeu-Olivares, 2009) rather than linear factor analyses. This meets the call of Chen who stated "one direction in future research is to systematically examine bias under various levels of invariance for categorical variables" (p. 1017).

HYPOTHESES

The primary objective of this study is to examine the impact of misspecified measurement parameters on structural relations in commonly used regression models. The impact is expected to depend on the role that the latent factor with non-invariant measurement part plays in the model – whether it is an independent or a dependent variable in structural relationships. The secondary objective is to examine whether the patterns of results for either role are similar across simple regression, regression with covariates, moderated regression, and mediation models. Based on previous research (e.g., Chen, 2008; Oberski, 2014), we hypothesize the following basic effects pertaining to misspecified factors.

Loading parameters

When factor loadings in the focal group are lower than in the referent group, and this is ignored, the variance of the latent factor in the focal group will be underestimated. The net effect will be an overestimation of the regression coefficient in the focal group when the mis-specified factor is the latent *X*-variable (independent, or predictor variable) in the structural model. Conversely, the net effect is an underestimation of the regression coefficient in the focal group when the misspecified factor is the latent *Y*-variable (dependent, or criterion variable). The effects will be reversed for the referent group.

Threshold parameters

When item thresholds in the focal group are lower (i.e., an acquiescent response style exists in the focal group: Cheung and Rensvold, 2000), and this is ignored, the latent factor mean in the focal group will be overestimated. While the effect on the mean is the strongest expected effect of the distorted factor metric, a distortion to the latent factor variance in the focal group is also expected, with the variance underestimated in the focal group. The net effect will be an overestimation of the regression coefficient in the focal group when the misspecified factor is the latent *X*-variable (independent, or predictor variable) in the structural model. Conversely, the net effect will be an underestimation of the regression coefficient in the focal

group when the misspecified factor is the *Y*-variable (dependent, or criterion variable). The effects are reversed for the referent group.

Loading and threshold non-invariance

When item thresholds are lower in the focal group (i.e., acquiescence is present) and factor loadings are also lower in the focal group the bias is expected be accentuated.

EXPERIMENTAL CONDITIONS

Modeling approach

Two approaches were compared; namely, (1) modeling measurement non-invariance, whereby non-invariant item parameters across groups were freely estimated, and (2) ignoring measurement non-invariance, whereby non-invariant item parameters were constrained equal across groups. Variances / residual variances of the latent factors were set to 1 in the referent group, and freely estimated in the focal group. Structural regression parameters were freely estimated in both groups.

Type of measurement non-invariance studied

Three types of item non-invariance were considered in the study. First, we examined the effect of factor loading (a.k.a. metric) non-invariance. Second, we examined the effect of threshold (a.k.a. scalar or strict) non-invariance in the form of an acquiescent response style. Finally, we considered the simultaneous effect of both the loadings and thresholds non-invariance.

Types of structural model

The first model considered is a simple regression model illustrated in **Figure 1**. Two regression coefficients, $\gamma_{11,1}$ and $\gamma_{11,2}$, quantify the paths linking the latent predictor variable to the latent criterion variable in the referent and the focal group, respectively.

Second, we considered a multiple regression model illustrated in **Figure 2**. In each group, the first regression coefficient represents the relationship between the predictor (target of our analysis) and the criterion, referred to as $\gamma_{11,1}$ and $\gamma_{11,2}$ in the referent and the focal group, respectively, while the second coefficient represents the relationship between the covariate and the criterion, referred to as $\gamma_{12,1}$ and $\gamma_{12,2}$. The population covariance of the predictors was fixed at zero, as it was not expected to impact results.

The third model examined was the mediated regression model illustrated in **Figure 3**. Four regression coefficients capture the structural relationships between variables. One coefficient per group, $\gamma_{11,1}$ and $\gamma_{11,2}$, link the predictor variables to the mediators in the referent and the focal group, respectively, and one coefficient per group, $\beta_{21,1}$ and $\beta_{21,2}$, link the mediators to the criterion variables.

Finally, we examined a moderated regression model illustrated in **Figure 4**. Two regression coefficients, $\gamma_{11,1}$ and $\gamma_{11,2,}$ quantifying the path linking the predictor and criterion latent variables in the referent and the focal group, respectively, summarize the variable relations here.

Test length and rating scale

We opted for a six-item measurement model for the target construct in our study. This is consistent with the test length reported by Meade and Lautenschlager (2004a,b) and Kim and Yoon (2011). We chose the polytomous items with three rating categories. This format is common in questionnaire research; for example, three response options (not true – somewhat true – certainly true) are used in the Strengths and Difficulties Questionnaire (Goodman, 1997), among many others. We used four indicators to model the auxiliary latent constructs in structural models, and this decision was not expected to impact the results of the analysis. Four item scales are often used in SEM research because four items is the minimum number of indicators required for a factor to be independently over-identified (Bollen, 1989). For these constructs, we opted for five-point Likert scales. Five point scales are often used due to the increased reliability that more scale points per item affords and is typical in personality questionnaire research (Furnham et al., 2013).

Proportion of non-equivalent items

We simulated four levels of invariance: zero non-invariant items (0%), one non-invariant item (16.67%), two non-invariant items (33.33%), and three non-invariant items (50%) out of six for measuring our target construct. With any greater non-invariance than this, researchers would likely be uncomfortable using the scale across subpopulations.

Sample size

We fixed sample size in all conditions at 1,000 respondents per group. The effect of sample size is out of scope for the present research, which focuses on model parameters and assumes that there is enough power in the study to estimate them. Limited information estimators that are required for speed in the context of models with categorical indicators are generally acknowledged to require larger sample sizes than linear factor models (Flora and Curran, 2004). Moreover, sample sizes of this magnitude are becoming more and more common in survey research due to advancing data collection technology.

Number of replications

A review of previous Monte Carlo research into measurement equivalence revealed that the number of replications ranged between a low of 50 replications per cell by Stark et al. (2006) and high of 500 replications per cell by Kim and Yoon (2011). We executed 1000 replications per cell, the highest number of replications of any of the studies reviewed.

Summary of experimental design

The Monte Carlo design involved 2 (modeling approaches)*3 (types of non-invariance)*4 (levels of non-invariance) *9 (types of structural models = 3 models where the latent variable assumed two structural positions, plus one model where the target latent variable assumed three structural positions) equals 216 conditions.

CREATING REPRESENTATIVE MODELS

Structural coefficient population values

We based the structural components of our models on an empirical study drawn from the applied literature. We searched for an example that contained the four different types of effects typically studied in psychological research, namely, simple regression coefficients, partial regression coefficients, mediated regression coefficients, and moderated regression coefficients. Wasti et al.

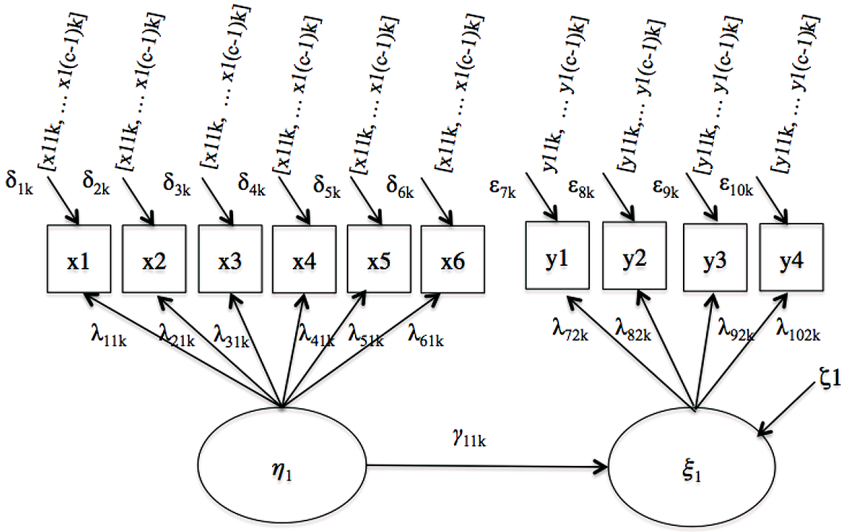


FIGURE 1 | Simple regression.

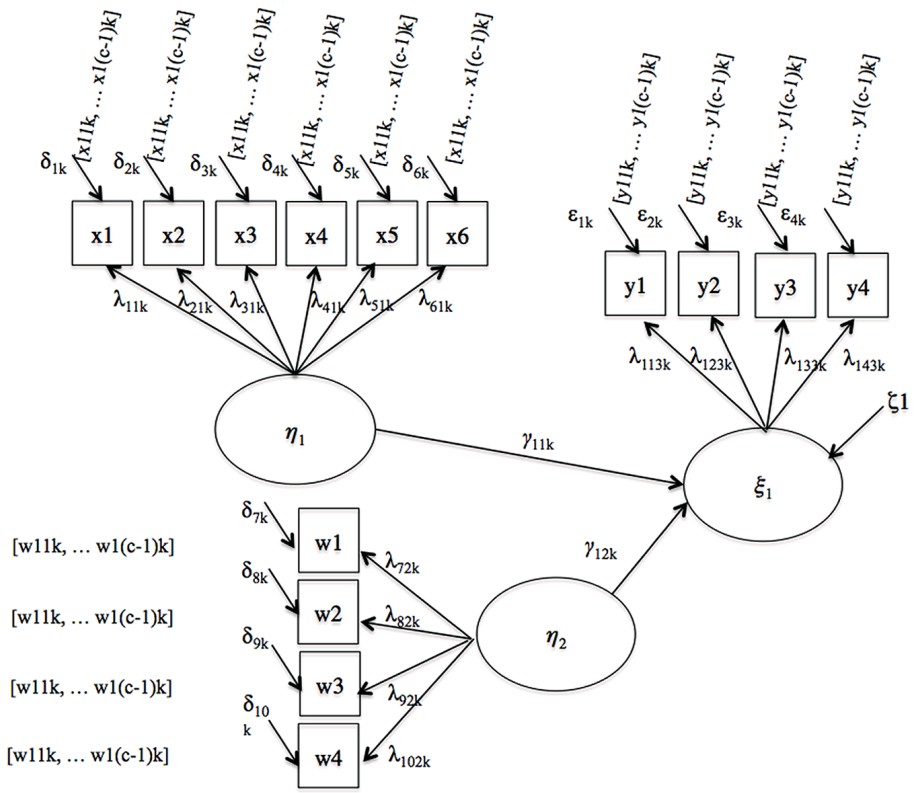


FIGURE 2 | Partial regression.

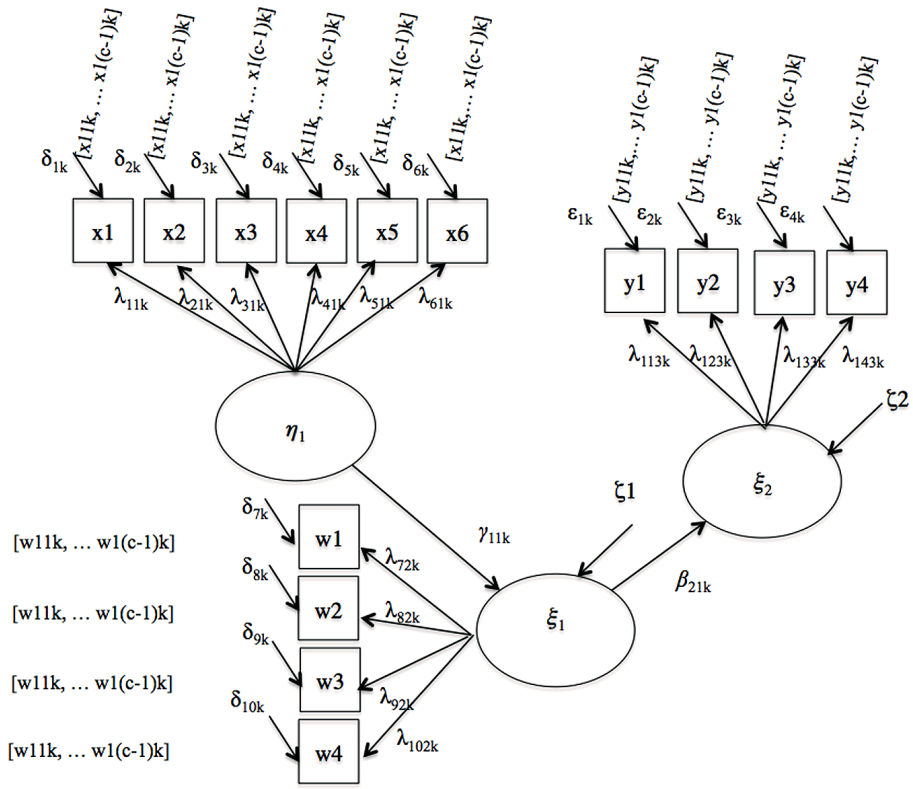


FIGURE 3 | Mediated regression.

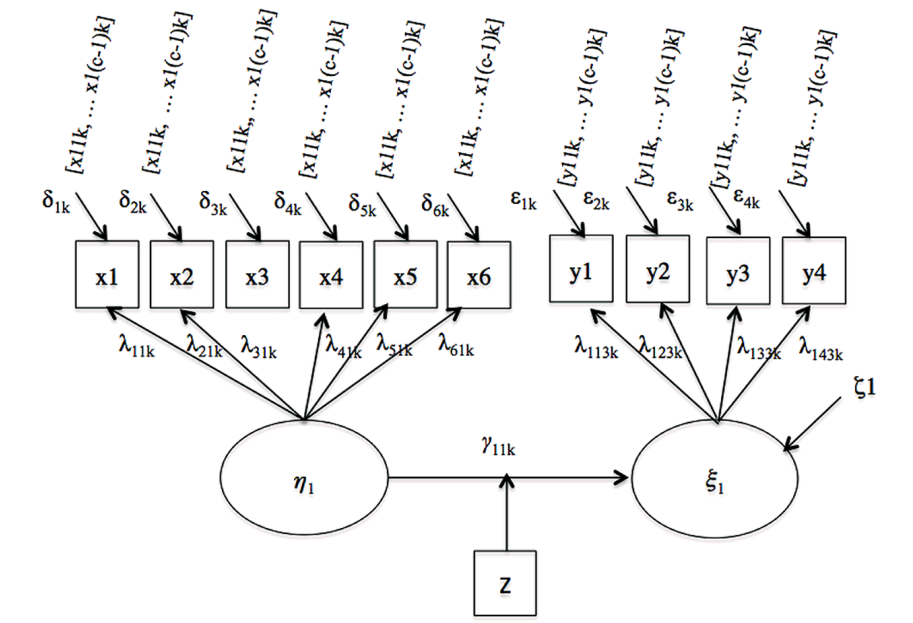


FIGURE 4 | Moderated regression.

(2000) presented a study with all four types of coefficient. These authors examined the structural equivalence of the relationship between the antecedents and consequences of sexual harassment between the United States and Turkey.

Sexual harassment was defined as an organizational cause of stress with performance related consequences. It was comprised of gender harassment (e.g., offensive, misogynist remarks), unwanted sexual attention, and sexual coercion. The relevant subsection of the model from Wasti et al. (2000) is shown in **Figure 5**. The model suggests that the two most important determinants of sexual harassment are job-gender context and organizational context. Job-gender context refers to how gender stereotyped the work is, while organizational context describes the features of the organization that communicate acceptance of sexual harassment. Sexual harassment, in turn, is negatively associated with job satisfaction.

We set the population structural regression values for all conditions based on paths from Wasti et al. (2000). The population simple regression effect was set at 0.22, the magnitude of the path between job-gender context and sexual harassment. The population partial regression coefficients were set at 0.22 and 0.37, which are the values of the paths between job-gender context and sexual harassment and organizational climate, respectively. The population mediation structural regression paths were set at 0.22 and −0.22, which correspond to the path from job-gender context to sexual harassment and from sexual harassment and job satisfaction. Wasti et al. (2000) also included the separate estimates of the relations between job-gender context and sexual harassment, at 0.47 for the United States sample and 0.18 for the Turkish sample. These values were used for moderated population structural regression values. The population variances for latent predictor variables were simulated equal to one, while population residual variances of latent criterion variables were simulated equal one minus the square of the structural coefficient.

Population factor loadings and simulated non-invariance

All simulated item loadings are presented in Appendix A. Target construct loadings were selected as representative of many questionnaire items using rating scales. These parameters are in the metric used with normal ogive IRT models, where the latent trait is scaled as having the mean of 0 and the variance of 1(i.e., the "theta" parameterization in Mplus; Muthén, 2013). Non-invariance was introduced by reducing focal group loadings on the second, fourth and sixth items by 50%. Our rationale for the 50% effect was that to be detectable in the structural relationship the impact of non-invariance needed to be at least moderate to strong, because existing empirical research suggests the effect of ignoring non-invariance on beta is small (Chen, 2008; Schmitt et al., 2011). For invariant indicators on auxiliary constructs, we followed Meade and Lautenschlager (2004a) by simulating loadings from a normal distribution with mean of 0.6 and a variance of 0.1. Error variances were set at one minus the square of the factor loadings. We then transformed these parameters to the IRT metric for the theta parameterization, by using formulas described by Wirth and Edwards (2007). While several authors have made a distinction between mixed and invariant patterns of loading differences across groups, Chen's (2008) literature review found just 7% of studies revealed mixed patterns of non-invariance. We focused on the so-called uniform pattern of invariance, the dominant outcome whereby all non-invariant items are lower in the focal group.

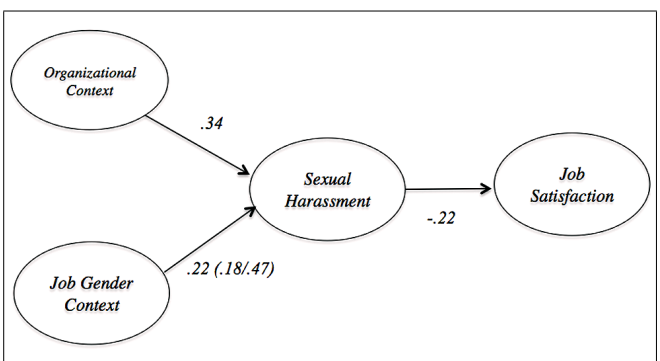


FIGURE 5 | Nomology of sexual harassment based on Wasti et al. (2000).

Population thresholds and simulated non-invariance

All simulated item thresholds are presented in Appendix A. Item thresholds of the target construct were representative of many questionnaire items using rating scales. We simulated non-equivalence on the second, fourth and sixth indicators on the focal construct by subtracting 0.8, 1.00, and 1.2 from the bottom threshold of the referent group. This did not disturb the relative ordering of the thresholds. For the auxiliary constructs, we followed Meade and Lautenschlager (2004b) to create thresholds by first drawing the lowest threshold from a normal distribution with mean of −1.7 and a standard deviation of 0.45. The remaining three thresholds were then created by adding constants to the lowest threshold for each item to give four thresholds per item. The constants were 1.2, 2.4, and 3.6. Threshold and loadings were transformed to the IRT metric using formulas from Wirth and Edwards (2007).

MATERIALS AND METHOD

ANALYSES

Model identification

To identify the metric of the latent factors, the loading of the first item of each factor was fixed at its population value in both groups. This allowed the latent variable variances and residual variances to be freely estimated in both groups. The means / intercepts of the latent factors were set to zero in the referent group and estimated freely in the focal group.

Specification of invariant and non-invariant conditions

Under the theta parameterization in Mplus the model where measurement invariance is imposed sees thresholds and loadings constrained equal across groups, error variances fixed at one in the referent group and free in the focal group, and factor means

fixed at zero in the referent group and freely estimated in the focal group. This setup was adopted in conditions where measurement non-invariance was ignored. The alternative models had the loadings and / or thresholds for the items known to be non-invariant freely estimated across groups.

Estimation

All models were fitted to polychoric correlations of the simulated item responses in MPlus 7.11 (Muthén and Muthén, 1998–2010) using the diagonally weighted least squares (DWLS) estimator with robust standard errors (denoted WLSMV in MPlus). Several simulation studies have shown that this estimator compares favorably to full information maximum likelihood (FIML) in comparable contexts to the current study (Muthén et al., Unpublished; Flora and Curran, 2004; Beauducel and Herzberg, 2006; Forero and Maydeu-Olivares, 2009). The simulations were executed by calling MPlus from the statistical computing environment R 3.0 using the package MPlusAutomation (Hallquist, 2011). All MPlus and R scripts are available at the following link: http://figshare.com/articles/Apples_Oranges_Monte_Carlo_Study/1060341.

Model performance

We examined five indicators of model performance. These included (1) the proportion of non-converged and inadmissible solutions; (2) how well the empirical chi-square distribution approximated the theoretical chi-square distribution; (3) the impact of the experimental conditions on power to reject the null hypothesis that the regression parameters were not significantly different from zero; (4) the relative bias, defined as the observed regression parameter minus the true parameter divided by the true parameter (Schunn and Wallach, 2005), and finally; (5) the coverage rate for all regression coefficients. Following Forero and Maydeu-Olivares (2009), we interpreted relative bias of less than 10% as acceptable, between 10 and 20% as substantial, and greater than 20% as unacceptable, and we considered coverage acceptable where the true parameter was captured by between 92.5 and 97.5% of 95% confidence intervals. We describe our results in the text; where boundary points are visible, we present results graphically.

RESULTS

MODEL ADMISSIBILITY AND GLOBAL FIT

Broad patterns observed following execution of the simulations greatly simplify presentation of results. First, all models converged to admissible solutions, indicating that the simulations ran well. We do not discuss model convergence further. The large sample size meant that power to detect whether regression parameters were significantly different from zero remained above 90% for all conditions of the study. We do not discuss power further. A clear difference emerged between the global fit results for models where increasing levels of non-invariance was modeled and where it was ignored. In all conditions where the non-invariance was modeled, the χ^2 test consistently approximated the theoretical chi-square distribution well at the first, second, fifth, and tenth percentiles. When the increasing level of non-invariance was ignored, χ^2 correctly rejected all models. We do not discuss χ^2 goodness of fit further.

We turn now to discuss relative bias and coverage of regression coefficients for all conditions. These results summarized in **Tables 1** and **2**. They are also graphically summarized and are available at http://dx.doi.org/10.6084/m9.figshare.1060341. However, we include a few typical graphical illustrations in the results sections that follow.

MODELED NON-INVARIANCE ON THRESHOLD, LOADING, AND COMBINED CONDITIONS

In the conditions where the measurement non-invariance was modeled, the regression parameters had acceptable coverage and relative bias. Based on this pattern, we further simplify reporting of results, describing coverage and relative bias only for models where the measurement non-invariance was ignored.

IGNORED NON-INVARIANCE OF THRESHOLDS

Under the threshold only non-invariance condition when non-invariance was ignored, relative bias always fell into the range defined as acceptable, i.e., less than 10% (one cell of the design showed 11%). When the non-invariance existed in the latent predictor, positive relative bias was observed in the focal group indicating over-estimation of the regression coefficient. When the latent variable with non-invariance occupied the mediator position, the path linking the predictor to the mediator in the focal group was over-estimated and the path linking the mediator to the ultimate criterion was underestimated. The opposite patterns of bias to those just described were observed in the referent group. The result for this condition is illustrated in **Figure 6**.

The effects of ignoring threshold only non-invariance on coverage, for the most part, parallel the results for relative bias. That is, there were minimal negative effects on parameter recovery for the regression coefficient. The coverage was acceptable in all but a small handful of conditions, i.e., with between 92.5 and 97.5% of 95% confidence intervals containing the true parameter. There was no discernible pattern in relation to whether the non-invariant construct occupied the position of the predictor or the criterion, or to whether the departure from acceptable coverage was on the target construct that exhibited bias or, in the case of the partial and mediated models, involved measurement invariant constructs. We thus conclude that structural coefficient coverage should not be a primary concern for researchers ignoring non-invariant item thresholds.

IGNORED NON-INVARIANCE OF LOADINGS

Simple regression with predictor non-invariance

Ignoring loading only non-invariant items led to acceptable negative relative bias for $\gamma_{11,1}$ (referent group) when one or two non-invariant loading were ignored, and substantial negative relative bias when three non-invariant loadings were ignored. Ignoring a single non-invariant item led to positive but acceptable relative bias for the regression parameter $\gamma_{11,2}$ (focal group). Bias for the focal group became positive and substantial for two non-invariant

Table 1 | Relative bias for ignored non-invariance conditions.

	Thresholds		Slopes		Thresholds and Slopes	
Items	$\gamma_{11,1}$	$\gamma_{11,2}$	$\gamma_{11,1}$	$\gamma_{11,2}$	$\gamma_{11,1}$	$\gamma_{11,2}$
Simple,predictor						
0	0.00	–0.01	0.00	0.01	0.00	0.01
1	–0.03	0.04	–0.05	0.06	–0.07	0.13
2	–0.04	0.09	–0.09	0.16	–0.12	0.35
3	–0.03	0.11	–0.11	0.23	–0.11	0.44
Simple,criterion						
0	0.00	0.01	0.01	0.01	0.01	0.00
1	0.03	–0.03	0.06	–0.06	0.09	–0.11
2	0.06	–0.06	0.13	–0.13	0.15	–0.25
3	0.06	–0.07	0.14	–0.18	0.13	–0.30

Items	$\gamma_{11,1}$	$\gamma_{12,1}$	$\gamma_{11,2}$	$\gamma_{12,2}$	$\gamma_{11,1}$	$\gamma_{12,1}$	$\gamma_{11,2}$	$\gamma_{12,2}$	$\gamma_{11,1}$	$\gamma_{12,1}$	$\gamma_{11,2}$	$\gamma_{12,2}$
Partial,predictor												
0	0.01	0.01	0.01	0.01	0.01	0.00	0.02	0.01	0.01	0.01	0.00	0.01
1	–0.03	0.00	0.05	0.01	–0.05	0.01	0.07	0.01	–0.08	0.00	0.13	0.01
2	–0.05	0.00	0.09	0.01	–0.09	0.01	0.17	0.01	–0.12	0.01	0.34	0.01
3	–0.04	0.01	0.10	0.00	–0.11	0.00	0.23	0.00	–0.13	0.00	0.44	0.01
Partial,criterion												
0	0.01	0.01	0.00	0.02	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.01
1	0.04	0.05	–0.03	–0.03	0.07	0.06	–0.05	–0.05	0.09	0.10	–0.10	–0.10
2	0.06	0.07	–0.05	–0.07	0.12	0.12	–0.12	–0.14	0.16	0.15	–0.24	–0.25
3	0.05	0.04	–0.06	–0.06	0.14	0.13	–0.19	–0.19	0.14	0.15	–0.29	–0.29

Items	$\gamma_{11,1}$	$\beta_{21,1}$	$\gamma_{11,2}$	$\beta_{21,2}$	$\gamma_{11,1}$	$\beta_{21,1}$	$\gamma_{11,2}$	$\beta_{21,2}$	$\gamma_{11,1}$	$\beta_{21,1}$	$\gamma_{11,2}$	$\beta_{21,2}$
Mediation,predictor												
0	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.02	0.02	0.02	0.01
1	–0.03	0.01	0.04	0.00	–0.05	0.01	0.07	0.01	–0.08	0.00	0.12	0.02
2	–0.04	0.01	0.09	0.01	–0.09	–0.09	0.18	0.01	–0.13	0.02	0.32	0.00
3	–0.03	0.01	0.09	0.00	–0.11	0.01	0.22	0.01	–0.11	0.03	0.43	0.01
Mediation,mediator												
0	0.00	0.02	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.02	0.00	0.01
1	0.04	–0.02	–0.03	0.04	0.05	–0.04	–0.06	0.05	0.10	–0.09	–0.09	0.12
2	0.05	–0.04	–0.07	0.09	0.10	–0.10	–0.13	0.17	0.16	–0.12	–0.26	0.33
3	0.06	–0.04	–0.07	0.10	0.14	–0.12	–0.19	0.23	0.13	–0.12	–0.31	0.45
Mediation,criterion												
0	0.01	0.02	0.00	0.01	0.01	0.02	0.00	0.01	0.02	0.01	0.00	0.01
1	0.01	0.05	–0.01	–0.02	0.01	0.07	0.01	–0.06	0.01	0.09	0.02	–0.10
2	0.01	0.07	0.00	–0.07	0.01	0.10	0.01	–0.14	0.00	0.15	0.00	–0.25
3	0.01	0.06	–0.01	–0.06	0.01	0.14	0.01	–0.18	0.02	0.14	0.01	–0.31

(Continued)

Table 1 | Continued

	Thresholds		Slopes		Thresholds and Slopes	
Items	$\gamma_{11,1}$	$\gamma_{11,2}$	$\gamma_{11,1}$	$\gamma_{11,2}$	$\gamma_{11,1}$	$\gamma_{11,2}$
Moderation,predictor						
0	0.01	0.01	0.01	0.00	0.02	0.00
1	–0.03	0.04	–0.04	0.06	–0.07	0.11
2	–0.04	0.08	–0.10	0.15	–0.14	0.30
3	–0.04	0.09	–0.13	0.20	–0.14	0.39
Moderation,criterion						
0	0.00	0.02	0.01	0.00	0.02	0.01
1	0.04	–0.03	0.07	–0.05	0.11	–0.10
2	0.08	–0.06	0.12	–0.13	0.17	–0.23
3	0.05	–0.07	0.16	–0.17	0.19	–0.27

*Regression parameter labels correspond to **Figures 1–4**; group 1 is the referent group, group 2 is the focal group.*

Table 2 | Coverage rates for ignored non-invariance.

	Thresholds		Slopes		Thresholds and Slopes	
Items	$\gamma_{11,1}$	$\gamma_{11,2}$	$\gamma_{11,1}$	$\gamma_{11,2}$	$\gamma_{11,1}$	$\gamma_{11,2}$
Simple,predictor						
0	0.95	0.96	0.95	0.95	0.95	0.95
1	0.94	0.95	0.93	0.94	0.92	0.93
2	0.93	0.96	0.90	0.94	0.86	0.82
3	0.92	0.94	0.89	0.90	0.88	0.75
Simple,criterion						
0	0.95	0.93	0.95	0.96	0.96	0.96
1	0.95	0.93	0.94	0.92	0.94	0.89
2	0.95	0.92	0.94	0.83	0.92	0.62
3	0.95	0.92	0.94	0.77	0.95	0.50

Items	$\gamma_{11,1}$	$\gamma_{12,1}$	$\gamma_{11,2}$	$\gamma_{12,2}$	$\gamma_{11,1}$	$\gamma_{12,1}$	$\gamma_{11,2}$	$\gamma_{12,2}$	$\gamma_{11,1}$	$\gamma_{12,1}$	$\gamma_{11,2}$	$\gamma_{12,2}$
Partial,predictor												
0	0.95	0.93	0.95	0.95	0.93	0.93	0.96	0.95	0.95	0.95	0.94	0.95
1	0.94	0.95	0.94	0.94	0.93	0.94	0.95	0.95	0.89	0.95	0.93	0.95
2	0.93	0.95	0.95	0.95	0.88	0.95	0.91	0.94	0.86	0.95	0.82	0.95
3	0.94	0.95	0.95	0.95	0.87	0.95	0.89	0.93	0.86	0.95	0.75	0.94
Partial,criterion												
0	0.95	0.95	0.94	0.95	0.95	0.93	0.94	0.95	0.95	0.94	0.95	0.95
1	0.94	0.95	0.94	0.93	0.95	0.93	0.94	0.95	0.95	0.94	0.89	0.84
2	0.96	0.95	0.92	0.88	0.95	0.93	0.94	0.95	0.91	0.91	0.65	0.42
3	0.94	0.94	0.93	0.89	0.95	0.93	0.94	0.95	0.93	0.92	0.50	0.28

(Continued)

Table 2 | Continued

	Thresholds				Slopes				Thresholds and Slopes			
Items	$\gamma_{11,1}$	$\beta_{21,1}$	$\gamma_{11,2}$	$\beta_{21,2}$	$\gamma_{11,1}$	$\beta_{21,1}$	$\gamma_{11,2}$	$\beta_{21,2}$	$\gamma_{11,1}$	$\beta_{21,1}$	$\gamma_{11,2}$	$\beta_{21,2}$
Mediation,predictor												
0	0.95	0.96	0.95	0.95	0.95	0.94	0.95	0.94	0.95	0.95	0.95	0.96
1	0.94	0.95	0.96	0.94	0.93	0.95	0.95	0.95	0.91	0.95	0.95	0.95
2	0.94	0.95	0.94	0.95	0.91	0.91	0.91	0.95	0.87	0.94	0.84	0.94
3	0.92	0.94	0.94	0.95	0.89	0.95	0.90	0.96	0.87	0.94	0.78	0.95
Mediation, mediator												
0	0.94	0.96	0.94	0.94	0.95	0.95	0.94	0.95	0.95	0.95	0.96	0.95
1	0.95	0.94	0.94	0.95	0.94	0.92	0.92	0.95	0.96	0.89	0.88	0.93
2	0.95	0.95	0.92	0.94	0.94	0.91	0.87	0.91	0.92	0.85	0.61	0.84
3	0.94	0.95	0.93	0.94	0.93	0.88	0.76	0.88	0.94	0.87	0.49	0.74
Mediation,criterion												
0	0.95	0.95	0.95	0.95	0.94	0.95	0.94	0.94	0.94	0.92	0.94	0.95
1	0.95	0.94	0.94	0.94	0.93	0.95	0.94	0.91	0.95	0.95	0.95	0.88
2	0.94	0.94	0.94	0.94	0.95	0.95	0.95	0.85	0.96	0.92	0.95	0.63
3	0.95	0.94	0.93	0.93	0.94	0.95	0.95	0.77	0.95	0.92	0.95	0.48

Items	$\gamma_{11,1}$	$\gamma_{11,2}$	$\gamma_{11,1}$	$\gamma_{11,2}$	$\gamma_{11,1}$	$\gamma_{11,2}$
Moderation,predictor						
0	0.95	0.96	0.94	0.93	0.95	0.94
1	0.94	0.94	0.94	0.95	0.92	0.92
2	0.94	0.93	0.91	0.88	0.87	0.62
3	0.94	0.93	0.87	0.79	0.87	0.44
Moderation,criterion						
0	0.95	0.95	0.94	0.95	0.94	0.95
1	0.96	0.92	0.94	0.88	0.94	0.79
2	0.96	0.89	0.94	0.74	0.94	0.38
3	0.96	0.83	0.93	0.55	0.93	0.18

loadings, and unacceptable for three non-invariant loadings. Coverage for $\gamma_{11,1}$ fell slightly below acceptable levels when two and three non-invariant items are ignored, and for $\gamma_{11,2}$, it fell slightly below acceptable when three non-invariant loadings were ignored.

Simple regression with criterion non-invariance

Relative bias for $\gamma_{11,1,}$ (referent group) was positive but acceptable for one ignored non-invariant loading and substantial and positive for two and three ignored non-invariant loadings. Non-invariant items led to negative relative bias for $\gamma_{11,2}$ (focal group) that was acceptable for a single item and substantial for two and three non-invariant items. Coverage for $\gamma_{11,1}$ was acceptable for all ignored non-invariance. Coverage for $\gamma_{11,2}$ dropped to an unacceptable level for even a single non-invariant item and progressively worsened with further ignored non-invariance.

Partial regression with predictor non-invariance

The referent group parameter $\gamma_{11,1}$ was characterized by negative but acceptable relative bias for one or two ignored non-invariant loadings, and substantial relative bias for three non-invariant items. The focal group parameter $\gamma_{11,2}$ showed acceptable positive relative bias for a single non-invariant item, reaching substantial and unacceptable levels of positive bias for two and three non-invariant items. Relative bias for $\gamma_{12,1}$ and $\gamma_{12,2}$ was acceptable across all levels of ignored non-invariance. Coverage for $\gamma_{12,1}$ and $\gamma_{12,1}$ remained acceptable for all levels of ignored non-invariance while coverage for both $\gamma_{12,2}$ and $\gamma_{12,2}$ fell just below acceptable levels when two or three non-invariant items were ignored.

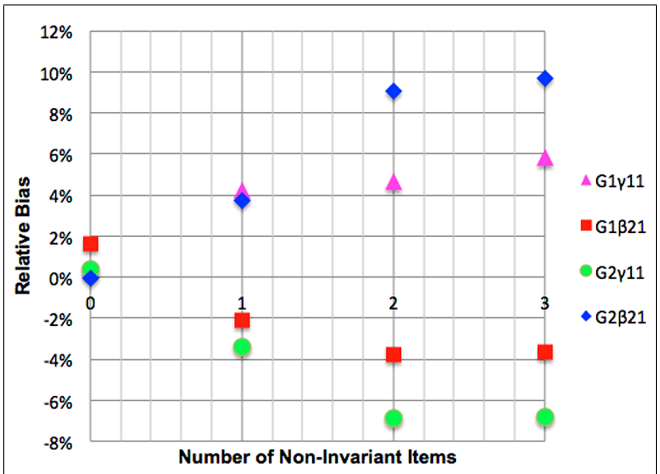


FIGURE 6 | Impact of mediator variable threshold non-invariance on relative bias of mediation betas.

Partial regression with criterion non-invariance

Relative bias for the referent group parameter $\gamma_{11,1}$ was positive and acceptable for one ignored non-invariant loading, and substantial for two and three ignored non-invariant loadings. The focal parameter $\gamma_{11,2}$ showed acceptable negative relative bias with one ignored non-invariant item, and substantial negative relative bias for two ignored non-invariant items, and unacceptable relative bias for three ignored non-invariant items. Relative bias for $\gamma_{12,1}$ was positive but acceptable for one ignored non-invariant item, and substantial for two or three ignored non-invariant items. Relative bias for $\gamma_{12,2}$ was negative and acceptable for one invariant item and substantial for two and three non-invariant items. Coverage for $\gamma_{11,1}$ and $\gamma_{12,1}$ was acceptable for all levels of ignored non-invariance. The same was true for $\gamma_{11,2}$ and $\gamma_{12,2}$.

Mediated regression with predictor non-invariance

Relative bias for the referent group parameter $\gamma_{11,1}$ was negative but acceptable for one and two ignored non-invariant loadings and substantial for three non-invariant loadings. Relative bias for the focal group parameter $\gamma_{11,2}$ was acceptable for a single non-invariant item, substantial for two non-invariant items and unacceptable for three non-invariant items. Relative bias of the coefficients $\beta_{21,1}$ and $\beta_{21,2}$ was acceptable for all levels of non-invariance. Coverage for all regression coefficients in referent and focal groups was acceptable except for $\gamma_{11,1}$ and $\gamma_{11,2}$ when two or three invariant items were ignored and $\beta_{21,1}$ and $\beta_{21,2}$ when two and three non-invariant items were ignored.

Mediated regression with mediator non-invariance

Relative bias for the referent group parameter $\gamma_{11,1}$ was positive but acceptable for one ignored non-invariant loading and substantial for two and three ignored non-invariant loadings. Relative bias for the focal group parameter $\gamma_{11,2}$ was negative and acceptable for a single non-invariant loading and substantial for two and three ignored non-invariant loadings. Relative bias of $\beta_{21,1}$ was negative and acceptable for one and substantial for two and three ignored non-invariant loadings, while relative bias for $\beta_{21,2}$ was positive and acceptable for a single non-invariant loading, substantial for two and unacceptable for three ignored non-invariant items. These results are presented graphically in **Figure 7**. Coverage rates for $\gamma_{11,1}$ were acceptable. Coverage for $\beta_{21,1}$ was acceptable for a single ignored loading but unacceptable for two and three ignored loadings. Coverage for $\gamma_{11,2}$ and $\beta_{21,2}$ reached unacceptable levels when two or three non-invariant items were ignored.

Mediated regression with criterion non-invariance

Relative bias for the referent group parameter $\gamma_{11,1}$ was positive and acceptable. Relative bias for the focal group parameter $\gamma_{11,2}$ was also positive and acceptable. Relative bias for $\beta_{21,1}$ was positive and acceptable for one ignored non-invariant loading, and substantial for two or three ignored non-invariant loadings. Relative bias for $\beta_{21,2}$ was negative and acceptable for one non-invariant item, becoming substantial for two and three non-invariant items. Coverage rates for all coefficients were acceptable except for $\beta_{21,2}$ where it became unacceptable when a single non-invariant items was ignored.

Moderated regression with predictor non-invariance

Relative bias for the referent group parameter $\gamma_{11,1}$ was positive but acceptable for one ignored non-invariant loadings and substantial for two and three ignored non-invariant items. Relative bias for the focal group parameter $\gamma_{11,2}$ was negative and acceptable for one ignored non-invariant item and substantial for two and three ignored non-invariant items. Coverage for $\gamma_{11,1}$ was acceptable for one and two ignored non-invariant items but unacceptable for three ignored non-invariant loadings, and the same was observed for $\gamma_{11,2}$.

Moderated regression with criterion non-invariance

Bias for the referent group parameter $\gamma_{11,1}$ was positive and acceptable for one ignored non-invariant loading and substantial for two and three items. Relative bias for the focal group parameter $\gamma_{11,2}$ was negative. It was acceptable for one and two

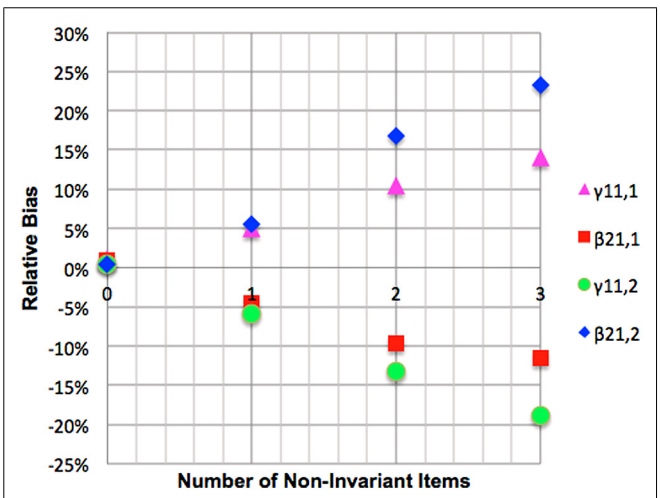


FIGURE 7 | Impact of mediator variable loading non-invariance on relative bias of mediation betas.

ignored non-invariant items and substantial for three items. Coverage for $\gamma_{11,1}$ became unacceptable when two non-invariant items were ignored. Similarly, the coverage for $\gamma_{11,2}$ fell to unacceptable levels when two or three non-invariant items were ignored.

THRESHOLD AND LOADING NON-INVARIANCE

Simple regression with predictor non-invariance

Non-invariant items caused negative relative bias for the referent group parameter $\gamma_{11,1}$. This bias was acceptable for one ignored non-invariant item and substantial for two and three ignored non-invariant items. Ignoring a single non-invariant item led to positive and substantial relative bias for the focal group regression parameter $\gamma_{11,2}$. This positive bias was unacceptable when two or three non-invariant items were ignored. Coverage for $\gamma_{11,1}$ was unacceptable for all levels of ignored non-invariance. Coverage for $\gamma_{11,2}$ deteriorated to unacceptable levels with two non-invariant items.

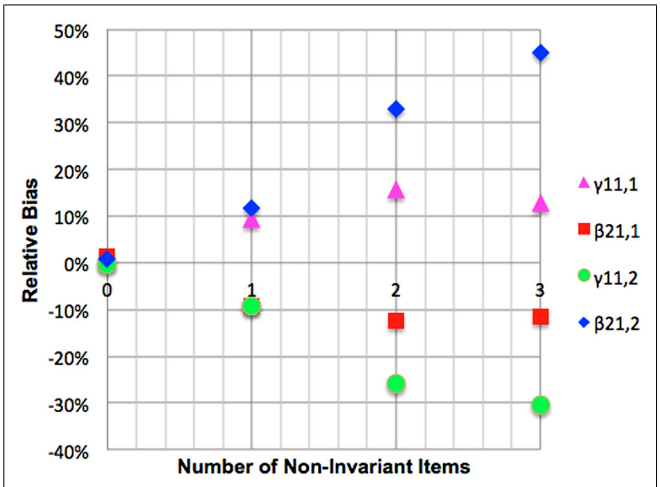


FIGURE 8 | Impact of mediator variable combined non-invariance on relative bias of mediation betas.

Simple regression with criterion non-invariance

Relative bias for $\gamma_{11,1}$ was positive and ranged from acceptable for a single non-invariant item to substantial for two and three non-invariant items. Non-invariant items led to negative relative bias for $\gamma_{11,2}$ that was substantial for a single item and unacceptable for two and three ignored non-invariant items. Coverage for the regression parameter $\gamma_{11,1}$ was acceptable for all ignored non-invariance except for two ignored items when it was marginally unacceptable. Coverage for $\gamma_{11,2}$, however, dropped to unacceptable levels as soon as a single non-invariant item is ignored. Coverage progressively worsened with further ignored non-invariance.

Partial regression with predictor non-invariance

Relative bias for $\gamma_{11,1}$ was negative and acceptable for a single item and substantial for two and three items. $\gamma_{11,2}$ suffered from substantial positive relative bias when even a single ignored non-invariant item was ignored. Relative bias increased to unacceptable levels for two and three non-invariant items. The relative bias of coefficients $\gamma_{12,1}$ and $\gamma_{12,2}$ was acceptable. Coverage for $\gamma_{12,1}$ and $\gamma_{12,2}$ remained acceptable for all levels of ignored non-invariance. However, the coverage rate for $\gamma_{11,1}$ was unacceptable for even one ignored non-invariant item, and coverage for $\gamma_{11,2}$ became unacceptable when two or three non-invariant items were ignored.

Partial regression with criterion non-invariance

Relative bias for $\gamma_{11,1}$ was positive and acceptable for one ignored item, becoming substantial when two or three items were ignored. $\gamma_{12,1}$ suffered from substantial negative relative bias with when one, two or three non-invariant items are ignored. Relative bias for $\gamma_{11,2}$ was negative and substantial for a single ignored item, and negative and unacceptable when two or three items were ignored. Relative bias for $\gamma_{12,2}$ was substantial when a single non-invariant item was ignored and unacceptable when two or three such items were ignored. Coverage for referent group parameters $\gamma_{11,1}$ and $\gamma_{12,1}$ became unacceptable when even a single non-invariant item was ignored and the coverage for $\gamma_{11,2}$ and $\gamma_{12,2}$ fell to unacceptable levels as soon as a two or three non-invariant items were ignored.

Mediated regression with predictor non-invariance

Relative bias for $\gamma_{11,1}$ was negative and acceptable for a single ignored non-invariant item and substantial for two or three ignored non-invariant items. Relative bias for this coefficient in the focal group, $\gamma_{11,2}$ was positive and substantial for a single ignored item and unacceptable when two or three non-invariant items are ignored. Relative bias of the coefficients $\beta_{21,1}$ and $\beta_{21,2}$ was near zero for all levels of non-invariance. Coverage for $\gamma_{11,1}$ fell below acceptable when one, two and three items are ignored, while the rate for $\gamma_{11,2}$ falls also falls below the acceptable threshold when two or three non-invariant items are ignored. Coverage for coefficients $\beta_{21,1}$ and $\beta_{21,2}$ remained acceptable for all levels of invariance.

Mediated regression with mediating non-invariance

Relative bias for $\gamma_{11,1}$ was positive and substantial for one, two or and three invariant items. Relative bias for $\gamma_{11,2}$ was negative and acceptable with one non-invariant item, worsening to unacceptable further non-invariance. Relative bias of $\beta_{21,1}$ was negative and acceptable for a single item and substantial for two or three items. Relative bias for $\beta_{21,2}$ was positive and substantial for one non-invariant item and this worsened with further non-invariance to unacceptable levels for two and three items. These results are presented graphically in **Figure 8**. Coverage rates for $\gamma_{11,1}$ with one ignored non-invariant item and unacceptable for two or three items. Coverage rates for $\gamma_{11,2}$ are unacceptable when even a single non-invariant item is ignored. Coverage rates for $\beta_{21,1}$ are also unacceptable when one or more non-invariant items are ignored while coverage for $\beta_{21,2}$ is unacceptable when two or three non-invariant item are ignored.

Mediated regression with ultimate criterion non-invariance

Relative bias for $\gamma_{11,1}$ was acceptable. Relative bias for $\gamma_{11,2}$ was also acceptable. Relative bias for $\beta_{21,1}$ was positive and acceptable when

one non-invariant item was ignored but positive and substantial when two or three non-invariant items are ignored. Relative bias for $\beta_{21,2}$ was negative and substantial with one non-invariant item and unacceptable for two or three non-invariant items. The coverage rate for all regression coefficients in this condition is acceptable except for $\beta_{21,2}$ where it deteriorates to unacceptable levels as soon as two non-invariant items are ignored and $\beta_{21,1}$ where coverage is unacceptable for one, two and three ignored non-invariant items.

Moderated regression with predictor non-invariance

When the non-invariant construct is the predictor, relative bias for $\gamma_{11,1}$ is negative but acceptable for a single item and substantial for two and three items. Relative bias for $\gamma_{11,2}$ is positive and substantial for a single ignored non-invariant item and becomes unacceptable for two and three ignored non-invariant items. Coverage for $\gamma_{11,1}$ and $\gamma_{11,2}$ became unacceptable as soon as a single non-invariant item was ignored and progressively worsened with increased ignored non-invariance.

Moderated regression with criterion non-invariance

Relative bias for $\gamma_{11,1}$ was positive and substantial for one to three ignored non-invariant items. Relative bias for $\gamma_{11,2}$ was negative and substantial as soon as a single non-invariant item is ignored becoming unacceptable when two or three items are ignored. Coverage for $\gamma_{11,1}$ is acceptable, while in the focal group the coverage for $\gamma_{11,2}$ falls away to unacceptable levels as soon as any non-invariance is ignored, worsening with further ignored non-invariance.

DISCUSSION

The issue of measurement invariance is important to any research or practice setting where the same measurement instrument is being used to assess individuals from different populations. Until now the focus of methodological work, looking at strategies for dealing with non-invariant measurement has mainly been restricted to measurement models and latent mean differences, with notable exceptions from Chen (2008) and Oberski (2014). The focus of the current article was on the implications of ignoring measurement non-invariance for accurate recovery of regression coefficients in full structural equation models. The results for the threshold conditions, loading conditions, and threshold and loading conditions showed that unacceptable relative bias and coverage were limited to the focal group regression parameter. While bias was observed for referent group parameters, this was never unacceptable. This pattern holds across simple, partial, mediated and moderated regression models. Under the conditions studied, i.e., lower focal group loadings, an acquiescent response style, or both, any path going into the non-invariant factor will yield an overestimated regression coefficient in the focal group, while any path coming out of the non-invariant factor will yield an underestimated regression coefficient in the focal group. The bias in the regression parameters emerged due to errors in the estimation of latent variances due to ignoring non-invariance. When predictor non-invariance is ignored due to lower focal loadings or acquiescent responding, focal regression coefficients are over-estimated (i.e., relative bias is positive) and when criterion non-invariance is ignored focal regression coefficients are under-estimated (i.e., relative bias is negative). When the non-invariant latent construct is in the mediator position, we see the path to it under-estimated and from it to the ultimate criterion variable overestimated. The aforementioned patterns were reversed in the referent group

IMPLICATIONS OF RESULTS REGARDING THRESHOLD NON-INVARIANCE

When either one or two items with non-invariant thresholds were not modeled, relative bias occurred in the aforementioned directions. However, ignoring the non-invariance led to relative bias below 10%, a level considered acceptable by Forero and Maydeu-Olivares (2009). Coverage rates were relatively unaffected. It is tempting for the applied researcher to conclude that they can ignore threshold non-invariance with impunity and argue that measures are consistent both across groups and past studies unless the number of non-invariant items is extreme. However, researchers must be careful to note that this is only the case if latent means are not a focus of the research (Steinmetz et al., 2009).

IMPLICATIONS OF RESULTS REGARDING LOADING NON-INVARIANCE

Across all types of structural models ignoring predictor non-invariance leads to over-estimation of focal regression coefficients, ignored criterion non-invariance leads to underestimation of focal regression coefficients, and when the non-invariant latent variable is in the mediating position the path to the mediator is underestimated while the path from the mediator in overestimated. The important difference between the loading only and threshold only conditions is that whereas relative bias never hit unacceptable levels in the threshold only condition, the relative bias in the loading condition routinely exceeded substantial and unacceptable thresholds when three non-invariant loadings were not modeled. Whereas coverage was not an issue for threshold only non-invariance, coverage became an issue in the loading only condition. The implications for the researcher are that ignoring non-invariance to permit scale comparability with previous research is okay for a single item with a non-invariant loading. However, when the non-invariance is on the loadings of two or more items and relational invariance is of critical importance, modeling the non-invariance is the best approach.

IMPLICATIONS OF RESULTS REGARDING LOADING AND THRESHOLD NON-INVARIANCE

Again we see consistency with the general pattern of the impact of ignored non-invariance on predictor latent variables leading to focal group positive relative bias, non-invariance on criterion latent variables leading to negative relative bias, and non-invariance on mediating latent variables producing mixed relative bias consistent with the role of the target variable (either independent or dependent). The main difference here is that the non-invariance causes problems for relative bias in the discussed directions at even lower levels of ignored non-invariance than for the thresholds only and slope only conditions. Serious problems are observed for relative bias when even one item with non-invariant loading and thresholds is not freely estimated across

groups. From this early stage the estimation accuracy of the regression parameters have unacceptable relative bias, a problem that worsens with further ignored non-invariance.

LIMITATIONS, FUTURE DIRECTIONS, AND CONCLUSION

This study only simulated data for three-point scales. While this number of scale points is regularly used in non-cognitive research, it is important that these results eventually be extended to dichotomous rating scales used in cognitive ability questionnaires, for example. This study also only simulated measurement invariance conditions for scales comprized of six items. While past Monte Carlo studies of measurement equivalence have also used six item scales, it would be beneficial to include the effect of ignoring measurement non-invariance on longer scales in terms of model recovery of regression parameters using alternate methods that are sometimes recommended to deal with longer scales in structural models (e.g., Yang et al., 2009).

We studied the impact of ignoring lower focal group loadings and focal group acquiescence. While Chen (2008) found lower focal group loadings were observed in over 90% of cases of measurement non-invariance, and acquiescence is a common response style, examining other conditions such as mixed loading non-invariance and extreme response styles are also important directions for future research. The current study also examined the impact of ignoring measurement non-invariance on regression parameter recovery assuming the distribution of the underlying latent variables is multivariate normal. It will be interesting to examine whether the results shown here generalize to conditions where this assumption is violated (c.f., DeMars, 2012). Finally, it is also important to examine the accuracy of Oberski's (2014) method in the context of regressions between factors indicated by categorical items.

Despite these limitations, the current study has important practical implications for researchers measuring constructs across multiple populations. The principal message from this study is that researchers must take the issue of measurement equivalence of the measures of latent variables seriously if they are interested in accurately estimating between construct relations using latent regression models. This is evident from the deteriorating trend in the accuracy of regression parameters as more non-invariance was introduced into the models. The current special issue and a rapidly expanding literature on measurement invariance both suggest that statisticians and psychometrics experts take the issue of measurement invariance extremely seriously. No doubt numerous applied researchers have caught themselves asking the question "does it really matter?" The short answer is to this question is "yes."

REFERENCES

Beauducel, A., and Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Struct. Equ. Modeling* 13, 186–203. doi: 10.1207/s15328007sem1302_2

Bentler, P. M., and Chou, C.-P. (1992). Some new covariance structure model improvement statistics. *Sociol. Methods Res.* 21, 259–282. doi: 10.1177/0049124192021002006

Beuckelaer, A., Lievens, F., and Swinnen, G. (2007). Measurement equivalence in the conduct of a global organizational survey across countries in six cultural regions. *J. Occup. Organ. Psychol.* 80, 575–600. doi: 10.1348/096317907X173421

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York: John Wiley & Sons, Inc.

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Struct. Equ. Modeling* 20, 518–540. doi: 10.1080/10705511.2013.797839

Borsboom, D., Romeijn, J. W., and Wicherts, J. M. (2008). Measurement invariance versus selection invariance: is fair selection possible? *Psychol. Methods* 13, 75–98. doi: 10.1037/1082-989X.13.2.75

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

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *J. Pers. Soc. Psychol.* 95, 1005–1018. doi: 10.1037/a0013193

Cheung, G. W., and Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *J. Cross Cult. Psychol.* 31, 187–212. doi: 10.1177/0022022100031002003

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., and Billiet, J. (2014). Measurement equivalence in cross-national research. *Annu. Rev. Sociol.* 40, 55–75. doi: 10.1146/annurev-soc-071913-043137

DeMars, C. E. (2012). A comparison of limited-information and full-information methods in mplus for estimating item response theory parameters for nonnormal populations. *Struct. Equ. Modeling* 19, 610–632. doi: 10.1080/10705511.2012.713272

Drasgow, F. (1982). Biased test items and differential validity. *Psychol. Bull.* 92, 526–531. doi: 10.1037/0033-2909.92.2.526

Drasgow, F. (1984). Scrutinizing psychological tests: measurement equivalence and equivalent relations with external variables are the central issues. *Psychol. Bull.* 95, 134–135. doi: 10.1037/0033-2909.95.1.134

Flora, D. B., and Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychol. Methods* 9, 466–491. doi: 10.1037/1082-989X.9.4.466

Forero, C. G., and Maydeu-Olivares, A. (2009). Estimation of IRT graded models for rating data: limited versus full information methods. *Psychol. Methods* 14, 275–299. doi: 10.1037/a0015825

French, B. F., and Finch, W. H. (2011). Model misspecification and invariance testing using confirmatory factor analytic procedures. *J. Exp. Educ.* 79, 404–428. doi: 10.1080/00220973.2010.517811

Furnham, A., Guenole, N., Levine, S. Z., and Chamorro-Premuzic, T. (2013). The NEO Personality Inventory–Revised: factor structure and gender invariance from exploratory structural equation modeling analyses in a high-stakes setting. *Assessment* 20, 14–23. doi: 10.1177/1073191112448213

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *J. Child Psychol. Psychiatry* 38, 581–586. doi: 10.1111/j.1469-7610.1997.tb01545.x

Hallquist, M. (2011). *MplusAutomation: Automating Mplus Model Estimation and Interpretation.* R package version 0.6. Available at: http://cran.r-project.org/web/packages/MplusAutomation/index.html

Holland, P. W., and Thayer, D. T. (1988). "Differential item performance and the Mantel-Haenszel procedure," in *Test Validity*, eds H. Wainer and H. Braum (Hillsdale, NJ: Erlbaum), 129–145.

Kankaraš, M., Vermunt, J. K., and Moors, G. (2011). Measurement equivalence of ordinal items: a comparison of factor analytic, item response theory, and latent class approaches. *Soc. Methods Res.* 40, 279–310. doi: 10.1177/0049124111405301

Kim, E. S., and Yoon, M. (2011). Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT. *Struct. Equ. Modeling* 18, 212–228. doi: 10.1080/10705511.2011.557337

Lubke, G. H., Dolan, C. V., Kelderman, H., and Mellenbergh, G. J. (2003). Weak measurement invariance with respect to unmeasured variables: an implication of strict factorial invariance. *Br. J. Math. Stat. Psychol.* 56, 231–248. doi: 10.1348/000711003770480020

Meade, A. W., and Lautenschlager, G. J. (2004a). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organ. Res. Methods* 7, 361–388. doi: 10.1177/1094428104268027

Meade, A. W., and Lautenschlager, G. J. (2004b). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Struct. Equ. Modeling* 11, 60–72. doi: 10.1207/S15328007SEM1101_5

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behav. Res.* 30, 577–605. doi: 10.1207/s15327906mbr3004_6

Millsap, R. E. (1998). Invariance in measurement and prediction: their relationship in the single-factor case. *Psychol. Methods* 2, 248–260. doi: 10.1037/1082-989X.2.3.248

Muthén, B. O. (2013). *IRT in Mplus.* Available at: http://www.statmodel.com/download/MplusIRT2.pdf

Muthén, L. K., and Muthén, B. O. (1998–2011). *Mplus User's Guide*, 6th Edn. Los Angeles, CA: Muthén & Muthén.

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Polit. Anal.* 22, 45–60. doi: 10.1093/pan/mpt014

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., and Chen, F. (2001). Monte Carlo experiments: design and implementation. *Struct. Equ. Modeling* 8, 287–312. doi: 10.1207/S15328007SEM0802_7

Saris, W. E., Satorra, A., and Sorbom, D. (1987). The detection and correction of specification errors in structural equation models. *Soc. Methodol.* 17, 105–129. doi: 10.2307/271030

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: a unified approach. *Psychometrika* 54, 131–151. doi: 10.1007/BF02294453

Schmitt, N., Golubovich, J., and Leong, F. T. (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, 412–427. doi: 10.1177/1073191110373223

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

Schunn, C. D., and Wallach, D. (2005). "Evaluating goodness-of-fit in comparison of models to data," in *Psychologie der Kognition: Reden and Vorträge anlässlich der Emeritierung von Werner Tack*, ed. W. Tack (Saarbrueken: University of Saarland Press), 115–154.

Stark, S., Chernyshenko, O. S., and Drasgow, F. (2006). Detecting DIF with CFA and IRT: toward a unified strategy. *J. Appl. Psychol.* 91, 1292–1306. doi: 10.1037/0021-9010.91.6.1292

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., and Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Qual. Quant.* 43, 599–616. doi: 10.1007/s11135-007-9143-x

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *Eur. J. Dev. Psychol.* 9, 486–492. doi: 10.1080/17405629.2012.686740

Wasti, S. A., Bergman, M. E., Glomb, T. M., and Drasgow, F. (2000). Test of the cross-cultural generalizability of a model of sexual harassment. *J. Appl. Psychol.* 85, 766–788. doi: 10.1037/0021-9010.85.5.766

Wirth, R. J., and Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychol. Methods* 12, 58–79. doi: 10.1037/1082-989X.12.1.58

Yang, C., Nay, S., and Hoyle, R. H. (2009). Three approaches to using lengthy ordinal scales in structural equation models: parceling, latent scoring, and shortening scales. *Appl. Psychol. Meas.* 34, 122–142. doi: 10.1177/0146621609338592

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 February 2014; accepted: 19 August 2014; published online: 17 September 2014.

Citation: Guenole N and Brown A (2014) The consequences of ignoring measurement invariance for path coefficients in structural equation models. Front. Psychol. ***5****:980. doi: 10.3389/fpsyg.2014.00980*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

APPENDIX A. MONTE CARLO MEASUREMENT MODEL PARAMETERS

Table A1 | Loadings.

	Target construct		Auxiliary construct
Item	**Referent**	**Focal**	**Both**
1	0.85	0.85	0.55
2	1.40	0.70	1.40
3	1.25	1.25	1.10
4	2.00	1.00	0.80
5	0.50	0.50	–
6	0.75	0.38	–

Table A2 | Thresholds.

	Target construct				Auxiliary construct			
	Referent		Focal		Both			
Item	***t*1**	***t*2**	***t*1**	***t*2**	***t*1**	***t*2**	***t*3**	***t*4**
1	–1.70	1.50	–1.70	1.50	–2.00	–0.80	0.20	1.70
2	–0.40	1.90	–1.20	1.90	–2.50	–0.90	0.30	1.90
3	0.70	2.30	0.70	2.30	–1.50	–0.20	0.80	2.30
4	–0.45	2.75	–1.45	2.75	–1.70	0.30	1.00	2.50
5	0.80	2.20	0.80	2.20	–	–	–	–
6	1.20	2.00	–0.20	2.00	–	–	–	–

frontiers in PSYCHOLOGY

METHODS ARTICLE
published: 05 February 2015
doi: 10.3389/fpsyg.2015.00087

Measurement equivalence in mixed mode surveys

Joop J. Hox[1]*, Edith D. De Leeuw[1] and Eva A. O. Zijlmans[2]

[1] *Department of Methodology and Statistics, Utrecht University, Utrecht, Netherlands*
[2] *Department of Methodology and Statistics, Tilburg University, Tilburg, Netherlands*

Edited by:
Alain De Beuckelaer, Radboud University Nijmegen, Netherlands

Reviewed by:
Jelte M. Wicherts, Tilburg University, Netherlands
Roel Schouteten, Radboud University Nijmegen, Netherlands

****Correspondence:***
Joop J. Hox, Department of Methodology and Statistics, Padualaan 14, 3584 CH Utrecht, Netherlands
e-mail: j.hox@uu.nl

Surveys increasingly use mixed mode data collection (e.g., combining face-to-face and web) because this controls costs and helps to maintain good response rates. However, a combination of different survey modes in one study, be it cross-sectional or longitudinal, can lead to different kinds of measurement errors. For example, respondents in a face-to-face survey or a web survey may interpret the same question differently, and might give a different answer, just because of the way the question is presented. This effect of survey mode on the question-answer process is called measurement mode effect. This study develops methodological and statistical tools to identify the existence and size of mode effects in a mixed mode survey. In addition, it assesses the size and importance of mode effects in measurement instruments using a specific mixed mode panel survey (Netherlands Kinship Panel Study). Most measurement instruments in the NKPS are multi-item scales, therefore confirmatory factor analysis (CFA) will be used as the main analysis tool, using propensity score methods to correct for selection effects. The results show that the NKPS scales by and large have measurement equivalence, but in most cases only partial measurement equivalence. Controlling for respondent differences on demographic variables, and on scale scores from the previous uni-mode measurement occasion, tends to improve measurement equivalence, but not for all scales. The discussion ends with a review of the implications of our results for analyses employing these scales.

Keywords: mixed mode survey, measurement equivalence, measurement invariance, mode effect, selection bias, propensity score adjustment

INTRODUCTION

Mixed mode surveys, which combine different modes of data collection, such as, face-to-face, telephone, and web, are becoming standard data collection tools (Biemer and Lyberg, 2003, p. 208; De Leeuw, 2005; De Leeuw et al., 2008; Dillman et al., 2014, p. 13). Mixed mode survey designs are attractive, because they are cost effective and because they can be successful in reaching different kinds of respondents (De Leeuw, 2005; Blyth, 2008). As a result, they have the potential to decrease both coverage errors and non-response errors, thereby increasing the representativeness of the final (combined) sample at affordable costs (Couper, 2011).

However, a combination of different modes in one survey, be it cross-sectional or longitudinal, can lead to different kinds of measurement errors (De Leeuw and Hox, 2011). An important distinction is in errors caused by the design and implementation of the survey and in mode inherent errors (De Leeuw, 2005; Dillman and Christian, 2005; Roberts, 2007). The former can be prevented; for instance, in the design phase survey questions are sometimes constructed differently for each mode (e.g., offering do-not-know in one mode but not in another). As a result respondents in particular modes are presented with different question formats, which will produce differences in responses. To avoid these question-format mode effects, Dillman et al. (2014, chapter 11) advocate the uni(fied)-mode design where equivalent questionnaires are developed for each mode in a mixed mode study. Mode effects can and should be reduced in the design phase as far as possible (see also De Leeuw et al., 2008 on designing for mixed-mode studies).

Mode inherent errors are part of the mode itself (Berzelak, 2014) and are not avoidable by clever design. A clear example is the way questions are presented to the respondent; this can be done visually or aurally; furthermore when questions are presented visually, the visual lay-out may convey extra information (e.g., Christian et al., 2007). As a consequence respondents in an interview survey may interpret the same question differently from respondents in an online survey and give a different answer, just because of the mode used. Another example is the presence or absence of an interviewer and its influence on sensitive questions (Dillman et al., 2014, chapter 8; Tourangeau et al., 2000, chapter 10).

We distinguish two different types of mode inherent effects on measurement (De Leeuw, 1992, chapter 7; Jäckle et al., 2010). First there are mode effects that only shift the response distribution; this produces differences in the mean or variance of scale scores between survey modes, but does not change correlations. The second mode effect is a change in the question-answer process and as a consequence the question is interpreted and answered differently. This can be the result of avoidable mode differences in wording, but also of mode inherent differences between aural and visual presentation. The latter has the potential

to produce measurements of constructs that are not equivalent between modes. In the worst case the instruments reflect qualitatively different constructs across modes. Both types of mode effect will be investigated in this study.

In addition to the measurement effect of survey mode on the response process, differential nonresponse across modes may play a role. Due to differential nonresponse, different types of respondents tend to end up in the different modes, even in randomized mode experiments. If these differences in sample composition across modes coexist with mode effects, this leads to confounding of substantive and methodological effects (Klausch et al., 2013; Vannieuwenhuyze and Loosveldt, 2013). For example, assume that we use a mixed mode web–interview survey to study drinking behavior. Web surveys attract younger respondents than traditional interviews (Couper, 2000; Mohorko et al., 2013). In addition, web surveys also elicit less socially desirable responses (e.g., Link and Mokdad, 2005). Since in our example the web mode is confounded with age, if the web respondents report more extreme drinking behavior we cannot distinguish whether the mixed mode data reveal a real relation between extreme drinking and age or if this is just the result of less socially desirable answers over the web.

Panel surveys pose their own challenge in this respect. Given the high costs of longitudinal panel surveys, there is a growing interest in applying mixed-mode data collection methods in such surveys (Dex and Gumy, 2011). Obviously, in longitudinal surveys that focus on measuring and explaining change, assessing and correcting mode effects is essential for a correct interpretation of trends over time. Compared to cross-sectional surveys, longitudinal surveys are in a special position. To assess selection effects, access is needed to auxiliary data not affected by mode effects. These data may come from elsewhere (e.g., a register), or the specific data are simply assumed to be unaffected by mode. Often the assumption is made that biographic information, such as, sex and age are measured without error. Even if that assumption is true, or if we have access to this information from a register, the problem remains that biographic variables usually are only weakly related to the substantive variables of interest and are therefore not very effective in assessing or correcting mode effects (Vannieuwenhuyze and Loosveldt, 2013). In longitudinal surveys that incorporate mixed mode data collection, preferably the first data collection occasion uses a single mode face-to-face interview, because this mode has the highest response rate compared to other or mixed modes (Hox and De Leeuw, 1994; Lozar Manfreda et al., 2008). The subsequent measurement occasions then shift to a less expensive mixed mode data collection. When this longitudinal survey design is followed, the first round of data collection provides a single mode data set that contains the substantive variables of interest measured with a constant mode effect. As a result, analysts have access to strong information to assess and correct mode effects.

Mode effects have been studied extensively for the traditional modes: face-to-face, telephone, and self-administered (e.g., mail) surveys not involving Internet. Most of these studies investigate simple mode effects, such as shifts in the response distributions of single questions, amount of missing data, or effects on sensitive questions. These studies typically find small differences, often indicating a dichotomy between survey modes with and without an interviewer (Groves, 1989; De Leeuw, 1992). When web surveys are added to the comparison, they tend to behave as self-administered paper-and-pen surveys. For an overview of such studies we refer to Christian et al. (2008), De Leeuw and Hox (2011), and Tourangeau et al. (2013, chapter 7).

Investigating measurement effects of data collection modes is difficult when individual questions are examined; repeated measures designs with several repeated measurement occasions are needed to distinguish between systematic and random measurement errors and true change over time. Alwin (2007) discusses the requirements for such designs, but also notes that their application to mode effect studies remains a challenge. When multi-item scales are involved, measurement equivalence can be investigated using models based on Item Response Theory (IRT) or Structural Equation Modeling (SEM). Since these models are closely related (Glockner-Rist and Hoijtink, 2003), we will only discuss measurement equivalence in mixed mode surveys using SEM. Given the potential confounding of selection effects by differential nonresponse in modes and by mode effects on measurement, we review only studies that also pay attention to differences in sample composition (i.e., selection effects) between the modes.

The question if measurement equivalence may be assumed, naturally occurs in cross-cultural comparisons across countries, where this is generally investigated in a Multigroup Confirmatory Factor Analysis (MCFA) using multigroup SEM. The assessment of measurement equivalence typically proceeds in steps (Jöreskog, 1971; Meredith, 1993; Vandenberg and Lance, 2000). The first step tests if the same factor model applies in different groups, traditionally countries, but in this particular study, modes are seen as groups. This is the weakest form of equivalence, configural equivalence, merely assuming that the different groups display the same pattern of factor loadings, i.e., the same number of factors, and these factors can be interpreted as similar because they have comparable loadings for their empirical indicators. The second step tests if (most of) these factor loadings can be constrained to be equal across all groups. If this holds we have (partial) *metric equivalence* (Vandenberg and Lance, 2000). When (partial) metric equivalence is achieved, one can validly test if the same structural model holds in all groups. The third step tests if (most of) the measurement intercepts can be constrained equal across all groups. If this holds we have (partial) *scalar equivalence* (Vandenberg and Lance, 2000). Full scalar equivalence is called strong measurement invariance in the psychometric literature (Meredith, 1993) and implies that the relationship between the observed score and the unobserved score on the latent factor of a person does not depend on group membership (Mellenbergh, 1989). Full scalar equivalence or strong measurement invariance allows variances and covariances between latent and observed scores to be different across groups. The psychometric literature also distinguishes strict measurement invariance, where the residual variances are also identical across groups (Millsap and Meredith, 2007). Since strict invariance is not necessary for valid comparisons across groups, we do not pursue strict invariance here.

When (partial) scalar equivalence is achieved, one can then investigate whether the latent means or actual sum scores differ

across the groups (step 4). For a valid comparison of groups it is not necessary that error variances be constrained equal (strict measurement equivalence), but if this constraint holds, this has the advantage that we are measuring with equal precision across groups. Regarding the minimal requirements for partial invariance, both Byrne et al. (1989) and Steenkamp and Baumgartner (1998) state that for each construct, in addition to the marker item that defines the scale –with marker item loading fixed at 1 and intercept fixed at 0–, at least one more indicator must have invariant loadings and intercepts across the groups. When the groups to be compared are different modes in a randomized mixed mode survey, the fourth step is extremely important. This fourth step tests if the latent mean or sum scores in different modes are equal. If not, we may have measurement equivalence, but the different modes still result in a response shift, with some modes reporting higher scores than other modes. This response shift points toward either a systematic bias in one of the modes or different systematic biases across modes.

What is known about measurement (in)equivalence across different modes? Probably the first mode experiment employing multigroup SEM is De Leeuw (1992, see also De Leeuw et al., 1996), who analyzed data from a national Dutch probability sample. They find non-equivalence, particularly between the mail survey mode on the one hand and the interviewer based face-to-face and telephone modes on the other. Although this study used random assignment to modes, there were small differences on age and gender, which were not controlled for in the SEM analyses.

Klausch et al. (2013) review empirical studies that evaluate measurement equivalence using MCFA following the sequence of steps outlined above. They report that comparisons of web and paper-and-pen surveys generally find full scalar equivalence and that measurement differences (i.e., nonequivalence) are more often found in comparisons of modes that do with modes that do not involve interviewers. However, most of the reviewed studies involve small samples of specific populations such as students or employees and not all of these studies control for potential selection effects. Below, we review in more detail studies that involve general populations and exert good control of selection effects.

Klausch et al. (2013) report a mode experiment in a crime victimization study using a random sample from the general population in The Netherlands. The respondents were randomly assigned to one of four modes: face-to-face, telephone, mail, and web; propensity scores based on eight socio-demographic variables were used to control for selection effects. The response categories formed either a three- or a five-point Likert scale. The data were analyzed with a MCFA specifying the variables as categorical and employing weighted least squares estimation. This approach involves estimation of thresholds for the observed variables, which allows an evaluation of the way respondents choose specific categories in the different modes. Klausch et al. (2013) report that interviewer-based surveys differ from self-administered surveys in measurement characteristics, with different systematic bias and different amounts of random error. The self-administered modes (i.e., mail and web) have lower category thresholds, indicating a greater tendency to agree to questions. Furthermore, the self-administered modes have lower error variances, which results in higher reliabilities for these modes.

Revilla (2013) compares data from two different large scale surveys in the Netherlands (the Dutch LISS internet panel and the Dutch contribution to the face-to-face ESS survey), both using large random samples from the general adult population. Using MCFA, she finds full scalar equivalence, including equal means on the latent variables, for four separate concepts. Although there is no explicit control for selection, Revilla (2013) reports that the two samples are very similar with respect to gender, age and education. Saris and Revilla (2013) analyze six Multi-Trait Multi-Method (MTMM) matrices from the same data sources. They focus on the *quality* of the responses, which they define as the strength of the relationship between the latent variable and the corresponding responses (Saris and Revilla, 2013, p. 2). They report finding few and small differences, if differences are found the questions in the LISS web survey have a higher quality than the corresponding questions in the face-to-face ESS survey.

Gordoni et al. (2012) investigate mode effects in a general survey of the Arab population in Israel, using face-to-face and telephone interviews. The survey topics concerned coexistence among the Arab minority in Israel, a topic that is potentially sensitive. For each survey mode an independent probability based sample was drawn. In addition, relevant demographic variables were included in the analysis as covariates. Gordoni et al. (2012) report full metric and partial scalar equivalence across the two modes. Measurement errors tended to be higher in the telephone mode than in the face-to-face mode.

Heerwegh and Loosveldt (2011) compare Likert scale responses in a national crime victimization study in Belgium. They use a mixed-mode design with telephone, mail and web modes. Assignment to modes was not random, but depended on the availability of a landline telephone number in the sampling frame. To control for differences between the modes, gender, age, education, job, and type of residence are included in the model as covariates. Conditional on these covariates, Heerwegh and Loosveldt (2011) report complete scalar equivalence between the combined mail/web and telephone modes. However, they do find a difference in the latent factor means: in the telephone mode the respondents show a more favorable attitude toward the police. Heerwegh and Loosveldt (2011) interpret these findings as the result of social desirability in the interviewer-based telephone survey.

Chang and Krosnick (2009) describe a national field experiment where the same questionnaire is administered to an RDD telephone sample, an Internet probability sample, and an Internet nonprobability, volunteer panel. After weighting all samples toward national demographics, they report that the two probability samples were more representative than the nonprobability sample, a difference that did not completely disappear after weighting. Compared to the probability based Internet sample, the telephone sample produced data that contained more random measurement error, more satisficing behavior, and more social desirability bias. These results were confirmed in a later laboratory study using students (Chang and Krosnick, 2010).

Summarizing: our review of large scale mode experiments that examine measurement equivalence across survey modes shows that all studies confirm configural measurement equivalence. This is not surprising, since all mode experiments investigated the

measurement equivalence of well-established scales, scales with proven reliability and validity. It would in fact be rather shocking if any mode would completely alter the structure of a reliable and valid scale that has been established in previous research. Many of the studies reviewed report full or at least partial scalar equivalence. When partial equivalence is found, the problems are more often situated in the intercepts or with ordinal measurement with the thresholds, which indicates just a shift in the response distributions across modes, and not in the factor loadings. Several studies report that error variances tend to be larger in interviewer based modes, especially in telephone surveys. It can be argued that the higher reliability in self-administered and especially internet modes simply reflects a common method effect, because in web and mail surveys several questions are usually presented together on one screen/page, instead of sequentially as is the case in interviews. This can enhance the intercorrelations between questions and thus increase the reliability, without increasing the validity. However, the studies of Saris and Revilla (2013) and Chang and Krosnick (2010) both suggest that the validity also increases.

Finally, *all* studies that report on demographics find small but systematic differences between the modes, even in randomized experiments. This finding confirms the importance of controlling for sample differences between modes in all survey mode experiments. In our study, we use propensity scores with covariate adjustment, a method that is suitable for controlling a potentially large number of covariates simultaneously. Our application of propensity score adjustment is described in detail in the next section of this paper.

The study reported here addresses three related research questions. The data source is a large longitudinal survey, which in its third wave of data collection changed over from single mode face-to-face to a mixed mode data collection. The first research question is whether the scales used do show measurement equivalence. If measurement inequivalence is found, this can be the effect of selection or of measurement differences due to mode. The second research question therefore is to what extent measurement equivalence improves if selection on demographic variables is controlled, and the third research question is to what extent measurement equivalence improves if scale scores from the earlier single mode data collections are added to the control variables.

DATA AND ANALYSIS METHODS

DATA

The data are from the Netherlands Kinship Panel Study (NKPS). The NKPS is a large-scale, nationally representative panel study on kinship in the Netherlands. Three waves of data collection have been conducted: wave 1 in 2002–2004, wave 2 in 2006–2007, and wave 3 in 2010–2011. Below we describe the data collection procedures briefly; full detail on design and fieldwork is available in the codebooks and questionnaires published on the NKPS homepage (www.nkps.nl), which also explains how researchers can obtain access to the NKPS data files. The NKPS data collection is funded by the Netherlands Organization for Scientific Research (NWO) and complies with standard NWO ethical requirements such as voluntary participation and informed consent.

The main NKPS wave 1 net sample consists of 8161 individuals who had responded to a face-to-face computer assisted interview (CAPI). Self-completion paper questionnaires were used to collect additional data from family members. In our analysis, we use only the data provided by the primary respondents, which are denoted as *anchor* in the NKPS files. In the second wave, a mixed mode design was introduced, where respondents were first approached for a face-to-face interview (CAPI), and computer assisted telephone interviewing (CATI) or computer assisted web interviewing (CAWI) were offered only at the end of the data collection period to sample members who had previously refused to participate or who had not been reached. This resulted in a net sample of 6091 individuals for the second wave. Very few respondents used the alternative options: about 3% used CATI and about 2% used CAWI. In our analysis, we have used only the face-to-face data from wave 2.

The NKPS wave 3 data collection was a fully sequential mixed mode design. The respondents were first offered to respond online by web mode (CAWI). CATI was offered at a later stage of data collection to sample members who had not responded to the web invitation. Next, CAPI was offered to those respondents who had not participated by Web or CATI. In the end, about 55% of the data was collected by web, 27% by telephone, and 18% by face-to-face interviews. The CAPI interviews employed show cards for some of the questions. The final response to the third wave of data collection was 4390 respondents. Since we are mainly interested in the mode effects in the third wave, we analyze the data of respondents who have responded to the third wave and also to the previous waves, leaving out nonrespondents on wave 2 or wave 3.

ANALYSIS METHODS

From the multi-item measures, we have selected 14 multi-item instruments that are assumed to be scales that measure a single underlying characteristic. Measuring an underlying characteristic by a scale has been referred to as *reflective measurement* (Bollen and Lennox, 1991). Some multi-item sets are not expected to form a scale, they are mere inventories of events or experiences that are expected to have an effect on respondents without reflecting an underlying characteristic. Such indices are referred to as formative measurement (Bollen and Lennox, 1991). Supplementary Material lists the multi-item scales used in our analysis.

When items are grouped in a scale, a Confirmatory Factor Analysis (CFA) can be used to check if they are indeed unidimensional and measure a single underlying characteristic. Multigroup CFA (MCFA) is then applied to evaluate their measurement equivalence across groups (research question 1). If the groups differ on some covariates, which indicates a selection process, various forms of adjustment are available. The correction approach used most often is conditioning on covariates that are related to the selection process and the target variables of the survey. Vannieuwenhuyze and Loosveldt (2013) call this approach *calibration*. The covariates are incorporated in the model by regressing the observed indicators on the covariates with equality constraints on the regression coefficients in all analyses, but allowing different intercepts (thresholds) across indicators and groups in the configural model. This follows the ANCOVA model described by Muthén (2002).

By regressing the observed indicators on the covariates, we assume that in the mixed mode design there is a selection on the observed variables. If the selection is on the latent variable instead, application of the Pearson–Lawley selection formulas leads to the conclusion that latent selection leads to an invariant factor model (Lawley and Maxwell, 1963; Meredith, 1964), even if the selection process is unknown. If the selection is on the observed indicators, application of the Pearson–Lawley selection formulae leads to the conclusion that the model is not invariant (Lawley and Maxwell, 1963; Muthén, 1989). We model the selection on the observed variables because in a general survey mostly demographics are available, which are expected to have only a small relationship with the latent variables. In a longitudinal survey we have access to measures from earlier measurement occasions, but we view these at best as proxies for the latent variables at later measurement occasions. By regressing the observed variables on the covariates, we expect that the factor model will change, in the direction of stronger measurement invariance.

When scale indicators have fewer than five categories we employ the ordered categorical variable methodology (Finney and DiStefano, 2006) as implemented in Mplus 7.1 (Muthén and Muthén, 1998-2012). For the measurement equivalence analysis, the consequence is that for categorical items the location parameter is no longer the intercept but a set of thresholds, which for scalar equivalence must be constrained equal across groups. Supplementary Material indicates which scales have categorical indicators, and provides reliability estimates (coefficient alpha) and some descriptive statistics for the scales.

If we have a large number of potential covariates, the covariate adjustment becomes unwieldy and also results in a complex model that estimates many regression coefficients for the covariates. To reduce the complexity of the model, that is, the number of covariates, propensity score methods can be used. Propensity scores were introduced by Rosenbaum and Rubin (1983) as a method to equalize an experimental and a control group on a set of covariates. The propensity score for a specific subject is the conditional probability of being assigned to treatment vs. control, given a set of covariates X. It can be viewed as a balancing score; a function $f(X)$ of the covariates, such that the conditional distribution of the set of covariates X given $f(X)$ is the same in both groups. The propensity score is used as a substitute for the entire set of covariates, thus considerably reducing the complexity of the model. Controlling for propensity scores can be performed by using them as covariates in an analysis (i.e., regression adjustment), or weights can be constructed based on the inverse of the propensity score (i.e., weighting adjustment). In our case we use regression adjustment. For a general overview of propensity score methods see Guo and Fraser (2010), for a review of propensity scores in surveys see Lee (2006).

Propensity scores are usually based on socio-demographic variables. This raises the question whether these are sufficient; propensity score methods assume that the propensity model includes all relevant variables. In longitudinal surveys, such as the NKPS, researchers have access to much richer information, namely the scores of respondents on the same variables collected on previous measurement occasions. For this reason we construct two different propensity scores: one based only on the socio-demographic variables and one based on the scales under investigation, measured in the previous wave that uses one single mode (face-to-face). Constructing a propensity score on the basis of observed sum scores in the first wave of data collection treats the scale scores as proxies for the latent variable scores at the first measurement occasion, which represents a stronger correction method than correcting on demographic information. Since having two sets of weights in one multivariate analysis is a complicated issue, we prefer applying the propensity score correction via regression adjustment. The first propensity score, based on demographics, is applied to answer research question two: "to what extent does measurement equivalence improve if selection on demographic variables is controlled for." The second propensity score, based on previously measured scale scores, is added to the covariate based on demographics, to answer the third research question: "to what extent does measurement equivalence improve if scale scores from the earlier single mode data collections are added to the control variables."

RESULTS

The Results section consists of two subsections. The first describes the construction of the propensity scores and the second presents the results of the measurement equivalence analyses.

CONSTRUCTION OF THE PROPENSITY SCORES

There are several methods to construct propensity scores, the most popular being logistic or probit regression (Guo and Fraser, 2010), which results in one optimal regression equation predicting group membership. The propensity scores are the regression based predicted probabilities of group membership, which can then be used as a single covariate or as a weighting variable. This works well in a two group context where an experimental and a control group must be balanced. In our case, there are three groups (the three modes CAPI, CATI, CAWI) and using multinomial logistic regression therefore produces always two regression equations, each contrasting one mode with the reference mode in the coding system. In order to establish if one optimal equation for each set of predictors may be sufficient to calculate a single propensity score, we decided to use discriminant analysis, as this has the potential to produce fewer relevant regression equations. In a discriminant analysis of three groups, a discriminant function is constructed, basically a regression function, that maximally discriminates between these three groups simultaneously. Next, a second discriminant function is constructed that maximally discriminates the three groups under the constraint that the second discriminant function is uncorrelated with the first discriminant function. Since the discriminant functions maximize discrimination between groups, successive discriminant functions decrease in importance, and it is usual to find fewer significant discriminant functions than there are degrees of freedom (the number of groups minus one). For a detailed description of discriminant analysis we refer to Tabachnick and Fidell (2013).

The first discriminant analysis used only the demographic variables gender, age, education, and urbanization, as measured in wave 1. Urbanization was not a significant predictor, and the final discriminant analysis is based on the demographic variables

gender, age and education. The first discriminant function captures 93.7% variance of the demographic variables, and a high score on this function reflects high age, being female and having a lower education. The second discriminant function explains 6.3% variance, and reflects being female with a high education. The high age, female and lower educated respondents represented by the first discriminant function are overrepresented in the CATI and CAPI modes and underrepresented in the CAWI mode in wave 3; the canonical correlation between this discriminant function and survey mode is 0.29. The female respondents with high education represented by the second discriminant function are underrepresented in CAPI, which indicates that they prefer to respond by telephone or web. Since the second discriminant function covers only 6% of the variance of the demographic variables, and since the associated canonical correlation with survey mode is only 0.08, it was decided to use only the first discriminant function as propensity score to correct for demographic differences. This propensity score is labeled D1 in the text and tables.

The second discriminant function is based on the scale scores in the second wave for those respondents who were interviewed using face-to-face. To avoid an accumulation of missing values in the scale scores, when some items of a scale were missing, scales were assigned the mean value on the available items. This was done after appropriate recoding for negatively worded questions and only if not more than 30% of the items of a scale were missing. If more items were missing, the scale score was assigned a missing value. A more serious missing value problem posed scales that apply only to a subsample of the respondents. For example, some scales enquire after parenting behavior, which of course only apply to respondents in certain age groups who actually have children. For all other respondents, such scales are assigned a missing value. Since SPSS Discriminant analysis uses listwise deletion to deal with missing values, simply specifying all available scales as potential predictors in a discriminant analysis would result in selecting only the small subgroup to which all of the scales apply. This not only dramatically reduces the number of respondents available for the analysis, but also restricts the analysis to a very specific subgroup of respondents. To avoid this, the discriminant analysis was carried out in a stepwise fashion. The first step includes as potential predictors all scales that apply to the entire sample, using forward selection to select only significant predictors. In the next steps, scales about partners and children were added. The scales about partners proved to be significant but the scales about children were not. Finally, a discriminant analysis was performed using all significant predictors. For the scales on partners, the missing values were imputed by the overall mean of the available values, and a dummy variable was added to indicate those cases where such imputation had taken place. As a result, the respondents to which these scales do not apply were not dropped from the analysis. The results of this discriminant analysis are summarized in **Table 1**.

The first discriminant function captures 89.1% variance of the wave 2 scales. A high score on this function reflects having a partner, feeling parental obligations, and division of homemaking tasks. The canonical correlation of this discriminant function with survey mode is 0.22. The second discriminant function explains 10.9% variance, and reflects having no partner combined with a low score on loneliness. Since the second discriminant function explains only 11% of the variance in the scales and the canonical correlation with survey mode is only 0.08, only the first discriminant function is used as propensity score. This propensity score is labeled D2 in the text and tables.

Table 1 | Standardized canonical discriminant function coefficients.

Scale	Function	
	1	2
Parental obligations	0.39	0.35
Parenthood	0.21	0.33
Loneliness	−0.20	−0.57
Conflicts partner	0.21	−0.00
Conflicts partner _missing	−0.79	0.43
Division homemaking tasks	0.30	−0.37
Division homemaking tasks _missing	0.07	0.08

Summarizing: the first propensity score D1 reflects differences in the sample composition of the three modes in demographic characteristics, and the second propensity score D2 reflects differences between the three modes in their scale scores on the previous, single mode, measurement occasion.

MEASUREMENT EQUIVALENCE TESTS

To simplify interpretation of the equivalence tests, the discriminant scores were standardized. The propensity scores were included in the measurement model by treating them as observed covariates; that is, regressing all observed indicators on the propensity scores, with equality constraints on the regression coefficients across the three modes (Muthén, 2002). Partial measurement models were investigated only if full equivalence did not hold and if the modification indices suggested that a partial equivalence model could improve the model fit. In **Table 2**, the qualification of the measurement equivalence includes partial equivalence. Decisions on model fit were done using the chi-square difference test (Jöreskog, 1971) because the models tested against each other are nested. In the case of categorical variables (<5 categories) the adjusted chi-square was applied using the DIFFTEST option in Mplus.

Table 2 indicates that full scalar measurement equivalence is rare for these scales. Correction for demographics (D1), or demographics plus wave two scales (D1+D2), in general improved the measurement equivalence. To explain the models behind the summaries in **Table 2**, we use (1) the Division Homemaking Tasks scale (scale 4), as example of a scale that has only configural equivalence; (2) the Parental Obligations scale (scale 9), as example of a scale where measurement equivalence clearly improves after propensity score correction; and (3) the Division of Childrearing Tasks scale (scale 6), as example of a scale that shows good measurement equivalence throughout.

CONFIGURAL EQUIVALENCE: THE DIVISION HOMEMAKING TASKS SCALE

The Division Homemaking Tasks scale showed only configural invariance, meaning that the same factor structure can be

Table 2 | Summary of results equivalence testing; (p) indicates partial equivalence.

Scale items (cat.) is categorical	Scale	No correction	Correction for D1	Correction for D1+D2
8A – 8E (cat.)	Support partner	Scalar (p)	Scalar (p)	–[a]
9A – 9E (cat.)	Conflicts partner	Scalar (p)	Scalar (p)	Scalar (p)
10A – 10D	Quality partner relationship	Scalar (p)	Scalar (p)	Scalar (p)
11A – 11E	Division homemaking tasks	Configural	Configural	Configural
13A – 13D (cat.)	Activities with children	No scale[b]	Configural	Configural
14A – 14D	Division childrearing tasks	Metric	Scalar	Scalar
24A – 24D	Family responsibility expectations	Scalar (p)	– [a]	Scalar (p)
24E – 24H	Filial responsibility expectations	Configural	Configural	Configural
24I – 24L	Parental obligations	Configural	Configural	Scalar
24M – 24P	Parenthood	Scalar (different means)	Scalar (different means)	Scalar (different means)
30A – 30D (cat.)	State vs. family	No scale[b]	No scale[b]	No scale[b]
32A – 32E	MHI-5	Metric	Configural	Metric
33A – 33K (cat.)	Loneliness	Metric (p)	Scalar (p)	Metric (p)
35M – 35P	Satisfaction with life	Scalar (p) (different means)	Scalar (p)	Metric (p)

The $\chi 2$ difference test for categorical analyses is computed using DIFFTEST.

[a] "–" Indicates that after imposing full scalar equivalence, the model did not fit adequately, but modification indices did not point to specific improvements to the model.

[b] RMSEA> 0.10 and CFI\TLI < 0.90.

imposed on these five items. The chi-square for the data with no correction is $\chi^2_{(15)} = 170.9$, $p < 0.001$, and values of the fit indices are RMSEA = 0.10 and CFI = 0.96. The model fit improved when corrections for selection effects were made. When we correct for demographics (D1) the chi-square is $\chi^2_{(23)} = 183.0$, $p < 0.001$ with RMSEA = 0.07 and CFI = 0.96. When both propensity scores (i.e., demographic D1 and previous wave scale scores D2) are used for correction, the chi-square for the configural equivalence model becomes $\chi^2_{(31)} = 189.4$, $p < 0.001$, with RMSEA = 0.06 and CFI = 0.96. Even with propensity score corrections, stronger levels of measurement equivalence than configural were not reached. **Table 3** presents the factor loadings and error variances for all data collection modes for the final configural equivalence model including the D1+D2 propensity score correction.

Although the data for this scale do not support either metric or scalar equivalence, it is clear that the loadings are nevertheless rather similar across the measurement modes. In fact, the correlation between any two columns of loadings is above 0.99. So it is tempting to invoke some kind of robustness and claim that modes can be combined and analyzed together, because the errors that are induced by this formally incorrect combination procedure are small and can be safely ignored. We come back to this in our discussion.

Table 3 | Factor loadings and intercepts Division Homemaking Tasks after D1+D2 propensity score correction: Configural equivalence.

Item	Loadings			Intercepts		
	CAPI	CATI	CAWI	CAPI	CATI	CAWI
11A	1.00	1.00	1.00	2.82	2.55	2.48
11B	0.66	0.65	0.78	2.83	2.68	2.58
11C	0.75	0.69	0.72	2.69	2.55	2.45
11D	−0.10	−0.05	−0.11	2.87	3.03	2.81
11E	−0.41	−0.47	−0.54	3.00	3.20	3.12

IMPROVEMENT WITH PROPENSITY SCORE CORRECTION: THE PARENTAL OBLIGATIONS SCALE

The Parental Obligations scale provides a nice example of improvement in measurement quality when the propensity score correction for selection is taken into account. Without adjustment, the chi-square for the configural equivalence model is $\chi^2_{(6)} = 25.5, p < 0.001$, and the fit indices are RMSEA = 0.05 and CFI = 1.00. Metric or scalar equivalence cannot be established. When we correct for demographics (D1) the chi-square for the configural equivalence model is $\chi^2_{(12)} = 38.8, p < 0.001$, and the fit indices are RMSEA = 0.04 and CFI = 1.00. Again, no metric or scalar equivalence can be established. With adjustment for both propensity scores (i.e., demographic D1 and previous wave scale scores D2) the chi-square for the strong scalar equivalence model becomes $\chi^2_{(32)} = 69.2, p < 0.001$ with fit indices RMSEA = 0.03 and CFI = 0.99. The fit indices are well within conventional limits for good fit, and we conclude that after D1+D2 correction full scalar equivalence is reached. To illustrate the effect of adding the correction for scale scores on the previous wave to the demographics, **Table 4** shows the factor loadings for the three modes in the configural model after D1 correction and under the heading *All* the common loadings in the final full scalar equivalence model (after D1+D2 correction).

In this example it is clear that using propensity score adjustment based on both demographics and previous wave scale scores leads to full scalar equivalence, which allows analyzing all data disregarding mode effects. It is interesting that without correction for the D2 propensity scores this is not the case. Again, we could argue that the loadings are very similar across the three modes, but in this case it is obviously better to use SEM analysis for

Table 4 | Factor loadings and intercepts error variances Parental Obligations Scale, for configural model after D1 correction and full scalar equivalence model after D1+D2 correction (All).

Item	Loadings				Intercepts			
	CAPI	CATI	CAWI	All	CAPI	CATI	CAWI	All
24I	1.00	1.00	1.00	1.00	2.41	2.48	2.53	2.49
24J	1.26	1.15	1.19	1.19	2.95	3.01	3.08	3.03
24K	1.02	1.02	1.18	1.14	2.66	2.69	2.81	2.74
24L	0.72	0.67	0.78	0.71	3.40	3.40	3.50	3.45

the substantive research questions, including the two propensity scores as covariates in all analyses.

FULL METRIC AND SCALAR EQUIVALENCE THROUGHOUT: THE DIVISION OF CHILDREARING TASKS SCALE

The Division of Childrearing Tasks scale shows full metric equivalence without correction, and reaches full scalar equivalence with either correction for only D1 (i.e., demographics) and for correction for both D1 and D2 (i.e., demographic plus previous wave scale) propensity scores. Without adjustment the chi-square for the metric equivalence model is $\chi^2_{(12)} = 15.2$, $p = 0.23$, and the fit indices are RMSEA = 0.02 and CFI = 1.00. After correction for demographics (D1) we have a scalar equivalence model with $\chi^2_{(26)} = 36.5$, $p = 0.08$, and the fit indices are RMSEA = 0.02 and CFI = 0.99. After correction for both demographics (D1) and previous wave scale scores (D2) this marginally improves into a scalar equivalence model with $\chi^2_{(32)} = 41.3$, $p = 0.13$, and the fit indices are RMSEA = 0.01 and CFI = 1.00.

Table 5 shows the loadings and intercepts of the models without correction (metric equivalence) and with correction for D1+D2 (full scalar equivalence). It is clear that adding covariates to the model brings the intercepts closer together, but from one model to the next the changes are very small.

CONCLUSION AND DISCUSSION

In this study, we addressed three related research questions. The first question is if the examined NKPS scales show measurement equivalence. The answer is that by and large they do, but in most cases we reach only partial measurement equivalence. The second research question is to what extent measurement equivalence improves if selection on demographic variables is controlled, and the third research question is to what extent measurement equivalence improves if scale scores from earlier, single mode, data collections are added to the control variables. In general, our analyses show that measurement equivalence improves if selection is controlled for, and that these measurement improvements improve more if in addition to demographics also previous wave scale scores are controlled for. Apparently, besides standard demographics, responses on an earlier wave play a role too. However, controlling for selection is not a panacea; there are a few cases where it does not improve the measurement equivalence at all, and one case (i.e., scale with items on activities with children) were adding the previous scale scores as covariate actually produces a weaker level of measurement equivalence.

One reviewer raised the question why correcting for propensity scores, which are a summary of demographic differences and scale score differences on the previous wave, only improves measurement equivalence in four out of 14 scales. One reason is that propensity score adjustment aims to correct for differential selection of respondents into specific modes. In addition to selection, our results point toward real mode effects in the measurement process. Berzelak (2014) makes a very useful distinction between mode inherent factors and context specific and implementation specific characteristics (see also De Leeuw and Berzelak, 2014). Mode inherent factors are given; examples are the involvement of interviewers in face-to-face and telephone surveys, absence of visual design elements in aural survey modes. Such factors are always present in specific modes. Context specific characteristics depend on social and cultural factors, such as familiarity with technology in the target population. These characteristics are difficult to influence, although they are likely to change over time. Implementation specific characteristics depend on the way a specific mode is actually implemented, such as the use of specific visual design elements in paper and web surveys. These are in principle under control of the researchers, and may be managed in a way to counteract context specific or mode inherent factors. The relatively small impact of our adjustment on the level of measurement equivalence suggests that mode inherent and context and implementation factors may be more important in mixed mode surveys than differential selection processes. If this is the case, research into mode effects and adjustment methods should attempt to include these characteristics, for example by collecting and using more paradata (Kreuter, 2015).

The results that we find depend of course on particularities of the instruments, data collection procedures, and sampling design employed in the NKPS. As large scale studies tend to make the switch from the expensive face-to-face mode to other modes, including mixed mode designs, other data will become available to investigate the generalizability of our results. In addition, it would be informative to carry out simulation research that manipulates potential selection mechanisms and employs different correction strategies.

The ideal situation is, of course, full scalar equivalence across modes. If full scalar equivalence is reached, we are justified in using scale sum scores in our analysis. If partial scalar equivalence is reached, such sum scores can be misleading, but scale means can be compared in structural equation models that include a partially equivalent measurement model. When only metric equivalence is reached, statements about differences in means, whether observed sum scores or factor means in a structural equation model, are not supported and cannot be validly made, but statements about covariances and correlations are still valid. When merely configural equivalence is reached, even statements about correlations can be invalid. In our analysis of the 14 NKPS scales, we find seven instances of (partial) scalar equivalence and three instances of (partial) metric equivalence. In three instances we find configural equivalence, and in one instance (state versus family support) the analysis shows that the items are not forming a scale according to any reasonable criterion.

If configural equivalence is established, we are measuring the same construct, but we measure it in slightly different ways in the

Table 5 | Loadings and intercepts division childrearing tasks scale.

Item	Uncorrected item scores				Correction D1		Correction D1 + D2	
	Loadings	Intercepts			Loadings	Intercepts	Loadings	Intercepts
		CAPI	CATI	CAWI				
14A	1.00	2.57	2.66	2.76	1.00	2.39	1.00	2.32
14B	1.21	2.63	2.48	2.67	1.27	2.45	1.27	2.34
14C	0.63	2.79	2.72	2.79	0.66	2.68	0.66	2.63
14D	1.10	2.74	2.56	2.70	1.13	2.48	1.11	2.37

different modes. If the actual values of the intercepts and loadings are close to each other across survey modes, as is the case in our example of the Division of Homemaking Tasks scale, it becomes very tempting to argue for some kind of robustness, even when metric or partly scalar equivalence does not hold. If the intercepts and loadings are very close, analysts might make a leap of faith, simply ignore any differences in intercepts and loadings, and work with SEM analyses of the combined data set or even compute sum scores for the scales and work with these, again on the entire data set. In our view, this may be defensible from a practical standpoint, but the burden of proof is on the researchers. They should make an attempt to estimate the amount of distortion produced by ignoring the real differences between intercepts and loadings across modes and demonstrate that the substantive effects they want to interpret are clearly larger than these measurement differences. Since analyses that follow this approach work by sweeping some real but hopefully small differences under the carpet, robust standard errors or bootstrapping should always be used to assess the real uncertainty in this case, since asymptotic statistical methods will underestimate the sampling variance.

A different way to deal with small measurement differences between survey modes is to employ a model that allows them and includes them explicitly in the model. Bayesian estimation is actually able to accomplish this, by introducing difference parameters in the model and by posing a prior distribution with a small variance for the difference parameters. For an example we refer to van de Schoot et al. (2013). This is a new and promising approach, but this is also an area that in our view needs more simulations and robustness studies to investigate when this approach works well and when it does not. We recommend that analysts that follow this approach carry out a sensitivity analysis to demonstrate that the specific choice of a prior does not have a large effect on the results for the substantive research questions.

There is a different approach to lack of measurement equivalence, which we have not explored in this study, because in our data the number of items in a scale was rather small (4-5). If there are enough items to form a scale there is always the option of dropping an item to improve the scale properties. The bare minimum to have a testable measurement model is four items for each latent variable and the bare minimum for testing measurement equivalence is three items (cf. Hair et al., 2010). Hence, if the number of items is larger than three or four, there is the option of finding the item that shows the largest amount of measurement non-equivalence and removing that particular item from the analysis. It follows that if the study is in a phase of developing measurement instruments and a mixed mode data collection is considered, it makes perfect sense to design measurement instruments with more than four or five items. From a SEM measurement point of view, this produces a number of potential superfluous items, that can in the analysis stage be sacrificed on the altar of measurement equivalence, and still leave a measurement model large enough that it can be tested.

A limitation in our discussion is that we have addressed mainly the issues that arise after the mixed mode data collection has been carried out. There is a large literature on designing questionnaires and fieldwork procedures that are aimed at minimizing mode effects by careful design. This is a broad topic, which is beyond the scope of this paper; for an extensive review of the issues that arise in designing mixed mode surveys we refer to De Leeuw et al. (2008) and Dillman et al. (2014).

Finally, we note that to distinguish between selection and mode measurement effects we need auxiliary information. In our analyses we used demographic data and data from a previous single-mode measurement occasion. Often the assumption is made that questions on factual demographic data are insensitive to mode measurement effects; in our case this information came from register data available from Statistics. Netherlands. Auxiliary information is also needed when attempts are made to adjust for mode measurement effects. Vannieuwenhuyze et al. (2011) discuss methods that use auxiliary data from a single-mode reference survey. Klausch et al. (in preparation) present a framework that uses a repeated single-mode survey on the same respondents, a design that in fact applies to panel surveys such as the NKPS where at least one measurement occasion is single-mode. De Leeuw (2005) and De Leeuw and Hox (2011) suggest to embed r real experimental design in the mixed-mode survey by assigning a subset of respondents at random to survey modes instead of allowing self-selection. All these approaches provide information needed to disentangle selection and measurement effect, which is a prerequisite to adjustment. Again, adjustment is a broad topic, and beyond the scope of this paper. However, it is important that when survey researchers design a mixed mode study, they anticipate the possible emergence of selection and measurement effects, and they must design the data collection in such a way that the necessary auxiliary information is made available.

ACKNOWLEDGMENTS

The Netherlands Kinship Panel Study is funded by grant 480-10-009 from the Major Investments Fund and by grant

481-08-008 from the Longitudinal Survey and Panel Funding of the Netherlands Organization for Scientific Research (NWO), and by the Netherlands Interdisciplinary Demographic Institute (NIDI), Utrecht University, the University of Amsterdam and the Erasmus University Rotterdam.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2015.00087/abstract

REFERENCES

Alwin, D. F. (2007). *Margins of Error: a Study of Reliability in Survey Measurement*. New York, NY: Wiley.

Berzelak, J. (2014). *Mode Effects in Web Surveys*. Unpublished Ph.D. thesis, University of Ljubjana, Slovenia. Available online at http://dk.fdv.uni-lj.si/doktorska_dela/pdfs/dr_berzelak-jernej.pdf (Accessed January 25, 2015)

Biemer, P. P., and Lyberg, L. E. (2003). *Introduction to Survey Quality*. New York, NY: Wiley.

Blyth, B. (2008). Mixed mode: the only 'fitness' regime? *Int. J. Market Res.* 50, 241–266.

Bollen, K. A., and Lennox, R. (1991). Conventional wisdom in measurement: a structural equation perspective. *Psychol. Bull.* 110, 305–314. doi: 10.1037/0033-2909.110.2.305

Byrne, B. M., Shavelson, R. J., and Muthén, B. O. (1989). Testing for the equivalence of factor and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Chang, L., and Krosnick, J. A. (2009). National surveys via RDD telephone interviewing versus the internet. Comparing sample representativeness and response quality. *Public Opin. Q.* 73, 641–678. doi: 10.1093/poq/nfp075

Chang, L., and Krosnick, J. A. (2010). Comparing oral interviewing with self-administered computerized questionnaires. an experiment. *Public Opin. Q.* 74, 154–167. doi: 10.1093/poq/nfp090

Christian, L. M., Dillman, D. A., and Smyth, J. D. (2007). Helping respondents get it right the first time: the influence of words, symbols, and graphics in web surveys. *Public Opin. Q.* 71, 113–125. doi: 10.1093/poq/nfl039

Christian, L. M., Dillman, D. A., and Smyth, J. D. (2008). "The effects of mode and format on answers to scalar questions in telephone and web surveys," in *Advances in Telephone Survey Methodology*, eds J. Lepkowski, C. Tucker, M. Brick, E. D. De Leeuw, L. Japec, P. Lavrakas, et al. (New York, NY: Wiley), 250–275.

Couper, M. P. (2000). Web surveys: a review of issues and approaches. *Public Opin. Q.* 64, 464–494. doi: 10.1086/318641

Couper, M. P. (2011). The future of modes of data collection. *Public Opin. Q.* 75, 889–908. doi: 10.1093/poq/nfr046

De Leeuw, E. D. (1992). *Data Quality in Mail, Telephone and Face to Face Surveys*. Amsterdam: TT-Publikaties. Available online at: http://edithl.home.xs4all.nl/pubs/disseddl.pdf (Accessed January 25, 2015)

De Leeuw, E. D. (2005). To mix or not to mix data collections in surveys. *J. Off. Stat.* 21, 233–255.

De Leeuw, E. D., and Berzelak, J. (2014). "Survey mode or survey modes?," in *The Sage Book of Survey Methodology*, eds C. Wolf, D. Joye, T. W. Smith, and Y.-C. Fu (Thousand Oaks, CA: Sage).

De Leeuw, E. D., and Hox, J. J. (2011). "Internet surveys as part of a mixed-mode design," in *Social and Behavioral Research and the Internet*, eds M. Das, P. Ester, and L. Kaczmirek (New York, NY: Routledge), 45–76.

De Leeuw, E. D., Hox, J. J., and Dillman, D. A. (2008). "Mixed-mode surveys: when and why," in *International Handbook of Survey Methodology*, eds E. D. De Leeuw, J. J. Hox, and D. A. Dillman (New York, NY: Erlbaum/Taylor and Francis), 299–316.

De Leeuw, E. D., Mellenbergh, G. J., and Hox, J. J. (1996). The influence of data collection method on structural models: a comparison of a mail, a telephone, and a face to face survey. *Soc. Methods Res.* 24, 443–472. doi: 10.1177/0049124196024004002

Dex, S., and Gumy, J. (2011). "On the experience and evidence about mixing modes of data collection in large-scale surveys where the web is used as one of the modes in data collection," in *National Centre for Research Methods Review Paper* (London: National Centre for Research Methods/Economic and Social Research Council). Available online at: https://kar.kent.ac.uk/39197/1/mixing_modes_of_data_collection_in_large_surveys.pdf (Accessed January 25, 2015)

Dillman, D. A., and Christian, L. M. (2005). Survey mode as a source of instability across surveys. *Field Methods* 17, 30–52. doi: 10.1177/1525822X04269550

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys*. New York, NY: Wiley.

Finney, S. J., and DiStefano, C. (2006). "Nonnormal and categorical data in structural equation modeling," in *Structural Equation Modeling. A Second Course*, eds G. R. Hancock and R. O. Mueller (Greenwich, CT: Information Age Publishing), 269–314.

Glockner-Rist, A., and Hoijtink, H. J. A. (2003). The best of both worlds: factor analysis of dichotomous data using item response theory and structural equation modeling. *Struct. Equ. Modeling* 10, 544–565. doi: 10.1207/S15328007SEM1004_4

Gordoni, G., Schmidt, P., and Gordoni, Y. (2012). Measurement invariance across face-to-face and telephone modes: the case of minority-status collectivistic-oriented groups. *Int. J. Public Opin. Res.* 24, 185–207. doi: 10.1093/ijpor/edq054

Groves, R. M. (1989). *Survey Errors and Survey Costs.* New York, NY: Wiley.

Guo, S., and Fraser, M. W. (2010). *Propensity Score Analysis. Statistical Methods and Applications*. Los Angeles, CA: Sage.

Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2010). *Multivariate Data Analysis*. London: Pearson.

Heerwegh, D., and Loosveldt, G. (2011). Assessing mode effects in a national crime victimization survey using structural equation models: social desirability bias and acquiescence. *J. Off. Stat.* 27, 49–63.

Hox, J. J., and De Leeuw, E. D. (1994). A comparison of nonresponse in mail, telephone, and face-to-face surveys. Applying multilevel modeling to meta-analysis. *Qual. Quant.* 329–344. doi: 10.1007/BF01097014

Jäckle, A., Roberts, C., and Lynn, P. (2010). Assessing the effect of data collection mode on measurement. *Int. Stat. Rev.* 78, 3–20. doi: 10.1111/j.1751-5823.2010.00102.x

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika* 36, 409–426. doi: 10.1007/BF02291366

Klausch, L. T., Hox, J. J., and Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Soc. Methods Res.* 42, 227–263. doi: 10.1177/0049124113500480

Kreuter, F. (2015). "The use of paradata," in *Improving Survey Methods: Lessons From Recent Research*, eds U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis (New York, NY: Routledge), 303–315.

Lawley, D. N., and Maxwell, A. E. (1963). *Factor Analysis as a Statistical Method*. London: Butterworths.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web surveys. *J. Off. Stat.* 22, 329–249.

Link, M. W., and Mokdad, A. H. (2005). Effects of survey mode on self-reports of adult alcohol consumption. A comparison of mail, web, and telephone approached. *J. Stud. Alcohol* 66, 239–245.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., and Vehovar, V. (2008). Web surveys versus other survey modes – A meta-analysis comparing response rates. *Int. J. Market Res.* 50, 79–104.

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Meredith, W. (1993).Measurement invariance, factor-analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Meredith, W. E. (1964). Notes on factorial invariance. *Psychometrika* 29, 177–185. doi: 10.1007/BF02289699

Millsap, R. E., and Meredith, W. (2007). "Factorial invariance: historical perspectives and new problems," in *Factor Analysis at 100: Historical Developments and Future Directions*, eds R. Cudeck and R. C. MacCallum (Mahwah, NJ: Erlbaum), 131–152.

Mohorko, A., De Leeuw, E., and Hox, J. (2013). Internet coverage and coverage bias in Europe: developments across countries and over time. *J. Off. Stat.* 29, 1–15. doi: 10.2478/jos-2013-0042

Muthén, B. O. (1989). Factor structure in groups selected on observed scores. *Br. J. Math. Stat. Psychol.* 42, 81–90. doi: 10.1111/j.2044-8317.1989.tb01116.x

Muthén, B. O. (2002). Beyond SEM: general latent variable modeling. *Behaviormetrika* 29, 81–117. doi: 10.2333/bhmk.29.81

Muthén, L. K., and Muthén, B. O. (1998-2012). *Mplus User's Guide, 7th Edition*. Los Angeles, CA: Muthén and Muthén.

Revilla, M. (2013). Measurement invariance and quality of composite scores in a face-to-face and a web survey. *Surv. Res. Methods* 7, 17–28.

Roberts, C. (2007). "Mixing modes of data collection in surveys: a methodological review," in *ESRC National Centre for Research Methods Briefing Paper*. Available online at: http://eprints.ncrm.ac.uk/418/1/MethodsReviewPaperNCRM-008.pdf

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41–55. doi: 10.1093/biomet/70.1.41

Saris, W. E., and Revilla, M. A. (2013). A comparison of the quality of questions in a face-to-face and a web survey. *Int. J. Public Opin. Res.* 25, 242–253. doi: 10.1093/ijpor/eds007

Steenkamp, J.-B. E. M., and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *J. Consum. Res.* 25, 78–90. doi: 10.1086/209528

Tabachnick, B. G., and Fidell, L. S. (2013). *Using Multivariate Statistics*. New York, NY: Pearson.

Tourangeau, R., Conrad, F. R., and Couper, M. P. (2013). *The Science of Web Surveys*. New York, NY: Oxford University Press.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). *The Psychology of Survey Response*. New York, NY: Cambridge University Press.

Vandenberg, R., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Org. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

Vannieuwenhuyze, J., and Loosveldt, G. (2013). Evaluating relative mode effects in mixed-mode surveys: three methods to disentangle selection and measurement effects. *Soc. Methods Res.* 42, 82–104. doi: 10.1177/0049124112464868

Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2011). A method for evaluating mode effects in mixed-mode surveys. *Public Opin. Q.* 74, 1027–1045. doi: 10.1093/poq/nfq059

Conflict of Interest Statement: The reviewer Jelte Wicherts declares that, despite being affiliated at the same department as the author Eva Zijlmans, the review process was handled objectively and no conflict of interest exists. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 March 2014; accepted: 15 January 2015; published online: 05 February 2015.

Citation: Hox JJ, De Leeuw ED and Zijlmans EAO (2015) Measurement equivalence in mixed mode surveys. Front. Psychol. ***6****:87. doi: 10.3389/fpsyg.2015.00087*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

METHODS ARTICLE
published: 25 July 2014
doi: 10.3389/fpsyg.2014.00745

Testing strong factorial invariance using three-level structural equation modeling

Suzanne Jak*

Department of Methods and Statistics, Faculty of Social Sciences, Utrecht University, Utrecht, Netherlands

Edited by:
Peter Schmidt, International Laboratory for Socio-Cultural Research, Russia

Reviewed by:
Jelte M. Wicherts, Tilburg University, Netherlands
Hermann Duelmer, University of Cologne, Germany

***Correspondence:**
Suzanne Jak, Department of Methods and Statistics, Faculty of Social Sciences, Utrecht University, Padualaan 14, 3584CH, PO Box 80.140, 3508TC Utrecht, Netherlands
e-mail: s.jak@uu.nl

Within structural equation modeling, the most prevalent model to investigate measurement bias is the multigroup model. Equal factor loadings and intercepts across groups in a multigroup model represent strong factorial invariance (absence of measurement bias) across groups. Although this approach is possible in principle, it is hardly practical when the number of groups is large or when the group size is relatively small. Jak et al. (2013) showed how strong factorial invariance across large numbers of groups can be tested in a multilevel structural equation modeling framework, by treating group as a random instead of a fixed variable. In the present study, this model is extended for use with three-level data. The proposed method is illustrated with an investigation of strong factorial invariance across 156 school classes and 50 schools in a Dutch dyscalculia test, using three-level structural equation modeling.

Keywords: measurement invariance, three-level structural equation modeling, cluster bias, measurement bias, multilevel SEM

INTRODUCTION

The purpose of this study is to show how three-level structural equation modeling (SEM) can be used to test for measurement invariance across the Level 2 and Level 3 clustering variables. The method is illustrated by testing measurement invariance across school classes and schools in a dyscalculia screening instrument.

MEASUREMENT INVARIANCE

In order to meaningfully compare test scores across groups, the test should be measurement invariant with respect to group membership. When a test is measurement invariant, the differences in test scores across groups can be attributed to differences in the constructs that were intended to be measured. The importance of measurement invariance is widely recognized (Mellenbergh, 1989; Millsap and Everson, 1991; Meredith, 1993; Vandenberg and Lance, 2000). In order to establish whether a test is measurement invariant across groups, one should test the equality of measurement parameters across groups. With continuous normally distributed test scores and continuous normally distributed latent variables (factors), the linear factor model is the suitable measurement model (Mellenbergh, 1994). If the relation between the factors and the test scores are equivalent across studies (i.e., if factor loadings are equal across groups), weak factorial invariance (also labeled as metric invariance) holds. If in addition the intercepts are equivalent across groups, strong factorial invariance (also labeled as scalar invariance) holds. With strong factorial invariance, the means of the factors can be meaningfully compared across the groups. If in addition the residual variances are equivalent (strict factorial invariance), the observed means can be compared across groups (Meredith, 1993; Widaman and Reise, 1997). In this study I focus on strong factorial invariance.

STRONG FACTORIAL INVARIANCE ACROSS MANY GROUPS

With a small number of groups, multigroup confirmatory factor analysis can be used to test the equality of measurement parameters (e.g., Wicherts and Dolan, 2010). If the number of groups is large, it may be convenient to view group as a random mode of variation, and use multilevel modeling (De Jong et al., 2007; Fox, 2010). See Muthén and Asparouhov (2013) for an overview of several fixed and random approaches to the study of measurement invariance across many groups.

Jak et al. (2013) showed how invariance restrictions across groups in a fixed model imply across level restrictions in a multilevel model. In a multilevel structural equation model, the covariance matrix is modeled as the sum of the covariance matrices at different levels (Muthén, 1990; Rabe-Hesketh et al., 2004). For a two-level model (for example, if the test scores are from students nested in school classes), the total covariance matrix can be decomposed in two independent covariance matrices:

$$\boldsymbol{\Sigma}_{\text{TOTAL}} = \boldsymbol{\Sigma}_{\text{LEVEL2}} + \boldsymbol{\Sigma}_{\text{LEVEL1}}. \quad (1)$$

The (pooled, within class) differences between students' scores are modeled by $\boldsymbol{\Sigma}_{\text{LEVEL1}}$. The average score of the school classes may also differ, these differences are modeled by $\boldsymbol{\Sigma}_{\text{LEVEL2}}$. At the different levels, distinct measurement models can be used to describe the covariances between the test scores. In this study we use linear factor models:

$$\boldsymbol{\Sigma}_{\text{LEVEL2}} = \boldsymbol{\Lambda}_{\text{LEVEL2}}\boldsymbol{\Phi}_{\text{LEVEL2}}\boldsymbol{\Lambda}^{\text{t}}_{\text{LEVEL2}} + \boldsymbol{\Theta}_{\text{LEVEL2}}, \quad (2)$$
$$\boldsymbol{\Sigma}_{\text{LEVEL1}} = \boldsymbol{\Lambda}_{\text{LEVEL1}}\boldsymbol{\Phi}_{\text{LEVEL1}}\boldsymbol{\Lambda}^{\text{t}}_{\text{LEVEL1}} + \boldsymbol{\Theta}_{\text{LEVEL1}}.$$

With p observed variables and k common factors, $\boldsymbol{\Phi}_{\text{LEVEL2}}$ and $\boldsymbol{\Phi}_{\text{LEVEL1}}$ are k by k covariance matrices of common factors,

$\boldsymbol{\Theta}_{\text{LEVEL2}}$ and $\boldsymbol{\Theta}_{\text{LEVEL1}}$ are p by p (diagonal) matrices with residual variances, and $\boldsymbol{\Lambda}_{\text{LEVEL2}}$ and $\boldsymbol{\Lambda}_{\text{LEVEL1}}$ are p by k matrices with factor loadings at Level 2 and Level 1, respectively.

METHODS

STRONG FACTORIAL INVARIANCE IN TWO-LEVEL MODELS

As explained by Jak et al. (2013), with two-level data, strong factorial invariance across clusters implies:

$$\boldsymbol{\Sigma}_{\text{LEVEL2}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}_{\text{LEVEL2}}\boldsymbol{\Lambda}^{\text{t}},$$

and

$$\boldsymbol{\Sigma}_{\text{LEVEL1}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}_{\text{LEVEL1}}\boldsymbol{\Lambda}^{\text{t}} + \boldsymbol{\Theta}_{\text{LEVEL1}}. \quad (3)$$

This means that if there is strong factorial invariance across clusters (so the factor loadings and intercepts are equal across school classes), the factor loadings are equal across levels, and there is no residual variance at Level 2 ($\boldsymbol{\Theta}_{\text{LEVEL2}} = \mathbf{0}$). All differences at the cluster (school class) level are thus differences in the common factor(s). If strong factorial invariance does not hold (i.e., if the intercepts differ across clusters), this results in residual variance at Level 2 ($\boldsymbol{\Theta}_{\text{LEVEL2}} \neq \mathbf{0}$). Strong factorial invariance across clusters can thus be investigated by testing the significance of Level 2 residual variance in a factor model with equal factor loadings across levels. This test is denoted the test for cluster bias. The cluster bias model can test whether strong invariance holds, but cannot differentiate between violations of weak and strong factorial invariance. The focus of this study is therefore on testing whether strong factorial invariance holds.

STRONG FACTORIAL INVARIANCE IN THREE LEVEL MODELS

With three-level data, such as test scores from students, nested in school classes, nested in schools, one may employ three-level structural equation modeling (Rabe-Hesketh et al., 2004). The total covariance matrix can be decomposed into three covariance matrices:

$$\boldsymbol{\Sigma}_{\text{TOTAL}} = \boldsymbol{\Sigma}_{\text{LEVEL3}} + \boldsymbol{\Sigma}_{\text{LEVEL2}} + \boldsymbol{\Sigma}_{\text{LEVEL1}}. \quad (4)$$

Here, $\boldsymbol{\Sigma}_{\text{LEVEL3}}$ refers to the covariance matrix of school averages, $\boldsymbol{\Sigma}_{\text{LEVEL2}}$ refers to the covariance matrix of class deviations from the school average, and $\boldsymbol{\Sigma}_{\text{LEVEL1}}$ is a covariance matrix of students deviations from the class average.

In a three-level factor model, the common factors also exist (have variance) at the third level. For example, with data from children in school classes in schools, the school averages in the test scores may be different. If strong factorial invariance across schools and across school classes holds, then the following model holds:

$$\boldsymbol{\Sigma}_{\text{LEVEL3}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}_{\text{LEVEL3}}\boldsymbol{\Lambda}^{\text{t}},$$

$$\boldsymbol{\Sigma}_{\text{LEVEL2}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}_{\text{LEVEL2}}\boldsymbol{\Lambda}^{\text{t}},$$

and

$$\boldsymbol{\Sigma}_{\text{LEVEL1}} = \boldsymbol{\Lambda}\boldsymbol{\Phi}_{\text{LEVEL1}}\boldsymbol{\Lambda}^{\text{t}} + \boldsymbol{\Theta}_{\text{LEVEL1}}, \quad (5)$$

Where $\boldsymbol{\Phi}_{\text{LEVEL3}}$ is a k by k covariance matrix of the common factors at Level 3. In this model, the common factor is the only source of variance at the class and at the school level (Rabe-Hesketh et al., 2004). If other variables than the common factor have influence at the school level, this will lead to residual variance at Level 3 ($\boldsymbol{\Theta}_{\text{LEVEL3}} \neq \mathbf{0}$), which means that measurement invariance across schools does not hold.

ILLUSTRATION

INTRODUCTION

Testing measurement invariance across in three-level models will be illustrated by testing strong factorial invariance across school classes and across schools in a dyscalculia screening test. Developmental dyscalculia is a learning difficulty specific to mathematics learning (Butterworth, 2005; Devine et al., 2013). Children with developmental dyscalculia have deficits in understanding basic concepts such as quantity conservation and reversibility, despite otherwise typically developing mental abilities (Kosc, 1974; Gross-Tsur et al., 1996). Dyscalculia is estimated to affect between 1.3 and 10% of the population, which is equivalent to the prevalence of dyslexia (Devine et al.). The screening of dyscalculia will often take place in the school, where a teacher administers the test to all children in the class. This way, the teacher can have influence on the test scores of the children. For example, one teacher may give better instructions than the other, leading to better test scores (less findings of dyscalculia) in the last school class. If this happens, the test is not measurement invariant across school class, as differences in test scores are not fully attributable to differences in dyscalculia (but to differences in quality of the instruction). At the school level, the school system may have influence on the test scores. For example, one school may have a curriculum that involves a different method to teach mathematics than another school. Or some schools may use more paper and pencil tests than other schools, leading to more experience of the students with a testing situation than others. If this is the case, two students that are equal in their levels of dyscalculia, may score differently on a screening test, depending on the school they are in. It is therefore important to establish measurement invariance of an instrument across school classes and schools. In this example, strong factorial invariance of a Dutch screening instrument for dyscalculia is tested across school classes and schools.

METHODS

Data

Respondents were 4527 students from 156 school classes in 50 schools in the Netherlands, of which 20 secondary schools and 30 primary schools. In all schools, the parent-teacher association or the teacher gave permission for the administration of the test. The test was administered by the teacher during regular school time. The students were in the first grade of the secondary school, or in the last 3 years of primary school. The schools were located across the country in a way that is representative of the distribution of people living in The Netherlands. For some schools, the class identifier was missing, in which case we treated all observations to be in one cluster. The average number of respondents per class was 29.02, the average number of respondents in each school

was 90.54. The mean age of the students was 11.42 ($SD = 1.27$), and 49.1% was a boy.

Instrument

The NDS (Nederlandse Dyscalculie Screener; Milikowski and Vermeire, 2013) is a screening instrument for dyscalculia. The screening instrument consists of eight subtests with a large set of items. For each subtest, the respondents try to answer as much items correctly as possible within 1 min. The score on each subtest is the amount of items answered correctly. The tests are long enough to ensure that no one can finish all questions in 1 min. See Appendix A for an overview of the content of the eight subtests. Before the respondents made the eight subtests they performed a control task, which does not involve numbers, to practice with the testing situation. The higher the score on each subtest, the lower the level of dyscalculia is assumed to be. As the scores are not recoded, the common factor that is assumed to underlie the test scores is actually the opposite of dyscalculia.

Analysis

All analyses were performed in the program M*plus* 7 (Muthén and Muthén, 1998–2012) using full information maximum likelihood estimation. In addition to the $\chi 2$ - statistic, the root mean squared error of approximation (RMSEA) and the comparative fit index (CFI) were used as measures of overall goodness-of-fit. RMSEA values smaller than 0.08 are satisfactory, values smaller than 0.05 indicate close fit (Browne and Cudeck, 1992). CFI values over 0.95 indicate reasonably good fit (Hu and Bentler, 1999).

First, the intraclass correlations and the significance of the variance at the class level and school level were inspected to decide whether multilevel modeling is actually necessary. Next, a measurement model is constructed at Level 1, with a saturated Level 2 and Level 3 model, so that all misfit stems from Level 1. Based on the final measurement model, a model with equal factor loadings across the three levels is fitted. Next, the significance of the Level 2 residual variance for all indicators is tested, by fixing all residual variance at Level 2 at zero. A significant chi-square difference in comparison with the free model indicates significant measurement bias across school classes. Finally, significance of Level 3 residual variance is tested by comparing the fit of a model with the residual variances at Level 3 fixed at zero with the model from the previous step. All tests are performed using a significance level of 5%.

Testing variances with the chi-square difference test in this way is not strictly correct Stoel et al. (2006). Correct testing requires the derivation of an asymptotic distribution of the likelihood ratio test statistic, which is a complex mixture of chi-square distributions. As this is beyond the scope of this work, I accept that the testing procedure is not correct, and keep in mind that it leads to an overly conservative test.

RESULTS

The intraclass correlations at the class level varied between 0.19 (Test 4) and 0.43 (Test 8), meaning that 19% to 48% of the variance in test scores is at the class level. At the school level the ICC's were much smaller, varying between 0.4% (Test 5) and 2% (Test 8). All variables showed significant variance at the class level, but not at the school level. Based on these results, one could decide to use two-level modeling instead of thee-level modeling. For the purpose of illustration, and because the interest is in differences between schools, I will continue the analyses using a three-level model.

First, the goal was to construct a measurement model at Level 1 with a saturated Level 2 and Level 3 model. Unfortunately, the model estimation did not converge when the Level 3 model was saturated, presumably because the saturated Level 3 model was overparameterized (i.e., some Level 3 correlations are actually zero). As a solution, the measurement model was specified with a saturated Level 2 model, and with corrections on the chi-square and standard errors to account for the dependency due to the school level (using "Type = Twolevel Complex" in M*plus*). A one-factor model fitted the data satisfactory according to the RMSEA, $\chi^2_{(20)} = 304.51$, $p < 0.05$, RMSEA $= 0.056$, CFI $= 0.93$. There was a modification index of a size 10 times larger than the others for the relation between Test 1 and Test 2. These tests are indeed quite similar (they both involve choosing the largest number, see Appendix A), so it seems to make sense that these tests share some specific variance. Adding a residual covariance between Test 1 and Test 2 leads to a better fitting model, $\chi^2_{(19)} = 135.69$, $p < 0.05$, RMSEA $= 0.037$, CFI $= 0.97$, with close fit according to the RMSEA and good fit based on the CFI. This model was accepted as the measurement model. Because it is not possible to model residual correlations at the higher levels in the next steps, the model was reparameterized by adding a factor on which Test 1 and Test 2 loaded. This factor was uncorrelated with the common factor, and both factor loadings are fixed at 1, so the model is equivalent with the model containing the correlated residuals (the estimate of the factor variance will be equal to the estimate of the residual covariance).

Using this measurement model, strong factorial invariance across school classes and schools is investigated. A model with equal factor loadings across levels fitted the data satisfactorily (see Model 1 in **Table 1**). Fixing the Level 2 residual variance at zero deteriorated the model fit significantly [$\Delta\chi^2_{(8)} = 2089.82$, $p < 0.05$], indicating that strong factorial invariance across school classes does not hold. Constraining the residual variance at Level 3 to be zero (and freely estimate Level 2 residual variance) did not lead to a significant deterioration of model fit, $\Delta\chi^2_{(8)} = 6.50$, $p = 0.59$. This indicates that strong factorial invariance across schools holds. The M*plus* syntax for the final model can be found in Appendix B.

Table 1 | Fit measures of the three-level models.

Model	*df*	χ^2	RMSEA	CFI
1. Baseline model (equal factor loadings across levels)	71	731.95	0.045	0.96
2. Strong factorial invariance at Level 2	79	2821.77	0.088	0.84
3. Strong factorial invariance at Level 3	79	738.45	0.043	0.96

Figure 1 shows the final model (Model 3) with unstandardized parameter estimates. By inspecting the significance of the residual variance for each indicator at Level 2, it appears that there is significant measurement bias across school classes for Test 1, Test 6, and Test 7. Using the parameter estimates, it can be calculated how much of the variance in these indicators is caused by class level variables other than (dys)calculia. The proportion of residual variance with respect to the total Level 2 variance is calculated as: Residual variance at Level 2/Total variance at Level 2. For Test 1 for example, the total variance at Level 2 is: $0.01 + 0.68^2 \times 0.56 + 0.02 = 0.28$, and the Residual variance at Level 2 is 0.02, so the proportion would be $0.02/0.28 = 0.071$. The proportion of

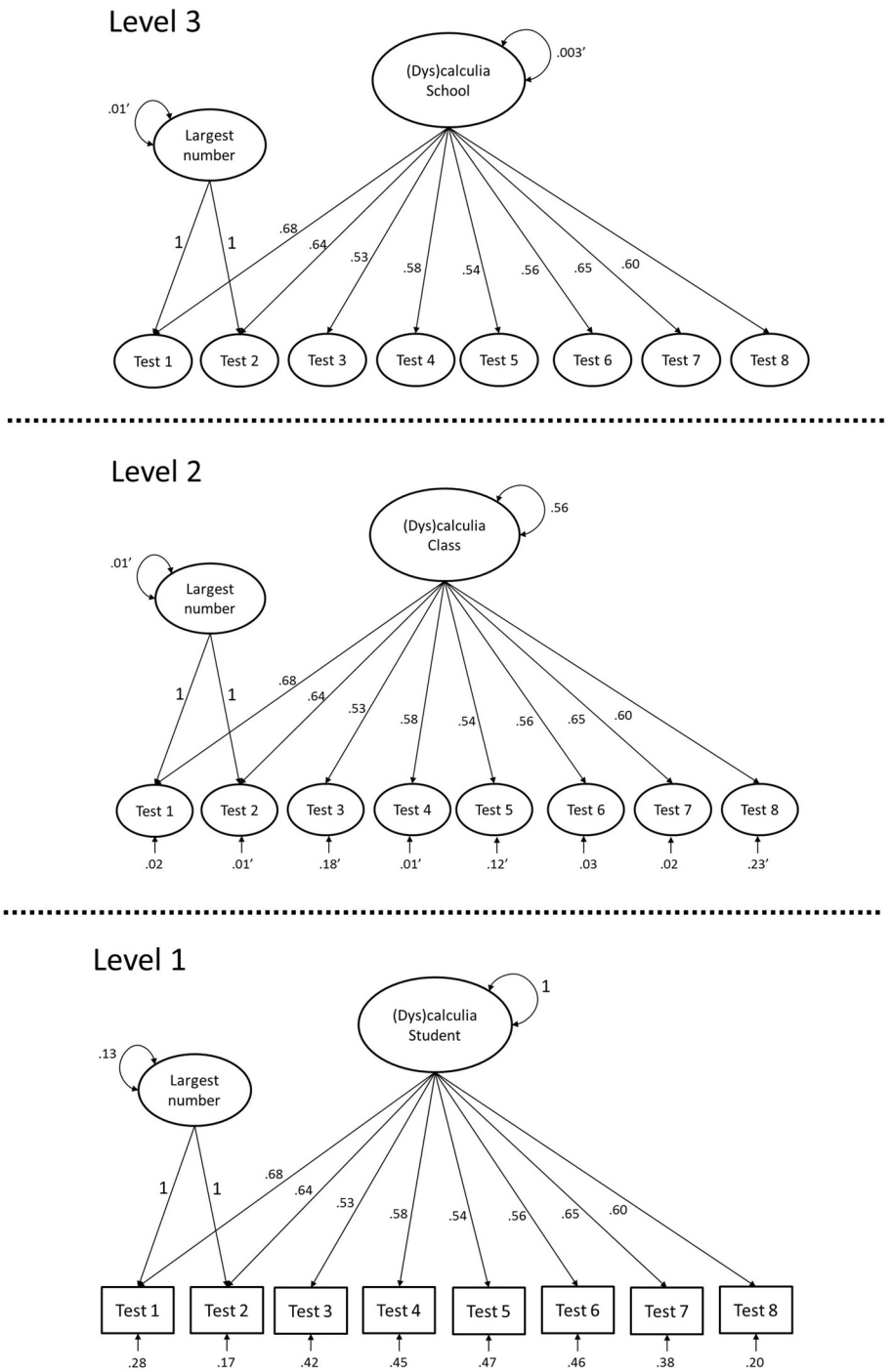


FIGURE 1 | A three-level factor model with equal factor loadings across levels and no residual variance at Level 3. Parameter estimates are unstandardized. Non-significance is indicated by an apostrophe (').

residual variance with respect to the total variance is calculated as: Residual variance at Level 2/Total variance at Level 1 + Level 2 + Level 3. **Table 2** gives an overview of these proportions for the three biased tests. Test 6 shows the most bias, followed by Test 7 and Test 1. However, the proportions of bias can be considered quite small in all tests.

Equality of factor loadings brings the factors on the same scale across levels, which means that the ICC of the factor can be calculated (Mehta and Neale, 2005; Kim et al., 2012). The ICC at Level 2 is equal to 0.56/ $(1 + 0.56 + 0.003) = 0.358$, indicating that 35.8% of the variance in dyscalculia is at the school class level. At the school level, the ICC is 0.003/ $(1 + 0.56 + 0.003) = 0.002$, so only 0.2% of the variance in dyscalculia is at the school level.

CONCLUSION

The analyses indicated that the screening instrument for dyscalculia cannot be considered fully measurement invariant across school classes. That is, in three of the eight subtests, differences across school classes cannot be fully attributed to differences in the average level of dyscalculia in the school classes. An explanation for the measurement bias can be found by looking at the content of the tests, and trying to distil the class level biasing factor. This is seldom easy, especially if the bias is small. In the current example, an explanation for class level bias in general could be the quality of the instruction that the teachers gave to the children. This is supported by the fact that Test 1 and Test 2 are quite similar (crossing out the largest number) and Test 7 and Test 8 are quite similar (subtraction and addition), but measurement bias across school classes is only found for the first tests of these pairs. In the second tests of each pair, the children already practiced with the type of assignment, rendering quality of the instruction less influential. Test number 6 is about filling in a number on a line, which can be viewed as a different from the other tests in that it forces respondents to visualize numbers on a straight line, which may not match the way students learn mathematics from their teacher. These is no cluster bias detected at the school level. As the number of schools, as well as the number of classes per school in this dataset are very small, a possible explanation of this non-finding is that the test for cluster bias did not have much power to detect bias at the school level.

DISCUSSION

In this study I illustrated how strong factorial invariance across the Level 2 and Level 3 clustering variable can be investigated. The employed method is only suitable to test strong factorial invariance, by rejecting models with zero residual variance at Level 2 or Level 3. However, the test cannot differentiate between violations of weak and strong factorial invariance. If $\boldsymbol{\Theta}_{\text{LEVEL2}} \neq \mathbf{0}$, this can also be caused by a difference in factor loadings across school classes, which is a violation of weak factorial invariance (Jak et al., 2013). So, if non-zero residual variance is detected, we know that strong factorial invariance does not hold, but we do not know if weak factorial invariance holds. An advantage of the current method is that factorial invariance with respect to Level 2 and Level 3 variables can be tested, even without having measured these variables. Non-zero residual variance at a level indicates bias with respect to some variable at that level, and can thus be viewed as a global test of measurement invariance with respect to any variable. If bias with respect to the clustering variable is found, covariates could be added to the model to explain the bias (Verhagen and Fox, 2012; Jak et al., 2014). In the current dataset this was not possible, as we did not have a measure of the supposed biasing factor, and other covariates at Level 2 did not have significant variance.

Table 2 | Proportions of variance caused by biasing variables at Level 2.

Test	Proportion bias Level 2	Proportion bias Total
Test 1	0.071	0.019
Test 6	0.146	0.031
Test 7	0.090	0.020

THE INTERPRETATION OF RESIDUAL VARIANCE IN MULTILEVEL MODELS

With equal factor loadings across levels, at the higher levels of a multilevel factor model, non-zero residual variance always represents measurement bias. This is not the case in single level data (or at Level 1), as we cannot distinguish variance caused by item specific factors from random measurement error variance.

In a factor model, residual variance stems from a residual factor (δ) that consists of two components, a structural component, s, and a random component, e (Bollen, 1998). With VAR() denoting variance:

$$\text{VAR}(\delta) = \text{VAR}(s) + \text{VAR}(e), \tag{6}$$

in which s represents a specific component, that is unique to the indicator, causing systematic variance in the test score. The remaining part of the residual variance is caused by a random component, e, representing measurement error. The expected value, denoted E(), of the structural component s may be non-zero, and could be interpreted as the intercept in a factor model:

$$\text{E}(s) = \tau. \tag{7}$$

The random component is unsystematic and has an expected value of zero:

$$\text{E}(e) = 0. \tag{8}$$

The residual variance of each indicator is thus equal to the sum of the variance of the two components, and the mean of the residual factor is equal to the mean of the structural component.

Zero structural residual variance represents invariance of the indicator with respect to all variables. As mentioned, in a single level model we cannot distinguish structural residual variance from measurement error variance, rendering it impossible to identify non-zero residual variance as measurement bias. At the second (and higher) level of a multilevel model, it *is* possible to test whether structural variance is present. Given that the cluster mean of the random component is expected to be zero (Equation 8), all residual variance at aggregated levels represents structural variance. Of course, if the number of observations per cluster is

very small, some random error variance may be aggregated to the higher level.

ALTERNATIVE APPROACHES WITH TWO-LEVEL DATA

The test for cluster bias is a useful addition to the existing set of structural equation modeling tools to investigate measurement bias. However, it is not the only test that can be used to investigate measurement invariance across clusters in multilevel data. One of the alternatives to is to test for measurement bias in a fixed effects model, i.e., in a multigroup model in which each cluster is a group. The equal factor loadings and intercepts across groups (clusters) in a multigroup model represent absence of cluster bias. Although this approach is possible in principle, it is hardly practical when the number of clusters is large. Muthén and Asparouhov (2013) describe an alternative way to circumvent the cumbersome strategy of multigroup modeling with large numbers of groups, using a 2-step procedure with Bayesian estimation. They introduce the concept of "approximate measurement invariance," referring to the analysis of measurement invariance across several groups using Bayesian SEM (BSEM), see also Van De Schoot et al. (2013). In Step 1 of the procedure (the analysis of approximate measurement invariance), in each group the difference between the group specific measurement parameter (factor loading or intercept) and the average of the particular parameter across all groups is estimated. The researcher can then identify the group with the largest difference between its measurement parameter and the average parameter as the most deviant group. In the next step, using BSEM, one estimates a model in which all factor loadings and intercepts are equal across groups, except for the groups that were identified as deviant in the previous step. This is similar to the use of modification indices with maximum likelihood estimation in a multigroup model, where the most deviant group will show the largest modification index in an analysis with equal factor loadings and intercepts. An advantage of the BSEM method is that it works well for the analysis of categorical variables, while maximum-likelihood estimation with categorical variables often leads to computational problems due to the numerical integration involved. A disadvantage of the approximate measurement invariance approach is that it relies on prior distributions for the model parameters, and different priors may yield different outcomes. Muthén and Asparouhov recommend zero-mean, small-variance priors for the difference parameters. However, the optimal size of the small-variance of the priors is a subject of debate. When trying to analyse the dyscalculia data using the BSEM method, it was unsuccessful due to the enormous computational load with 156 groups. Indeed, I have not seen applications of the BSEM method with large numbers of groups.

A framework for the detection of measurement bias across large numbers of groups within Bayesian Item Response Theory (IRT) is given by Verhagen and Fox (2012), using multilevel random item effects models (De Jong et al., 2007; Fox and Verhagen, 2010). Verhagen and Fox estimate a random effects parameter for all measurement parameters in the model (i.e., discrimination parameters and difficulty parameters in an IRT model), and test which of the measurement parameters have significant variance across clusters using Bayes factors or using the Deviance Information Criterion (DIC). Consequently, the cluster level variance in item parameters may be explained by adding covariates to the model. The approach of Verhagen en Fox is similar to the approach in this article in some respects. Both approaches treat groups as randomly drawn from a population of groups. Both approaches test the hypothesis of zero variance of parameters at the cluster level, and both allow for the explanation of non-zero variance by cluster level variables. The main differences between the two approaches relate to the modeling framework (multilevel IRT vs. multilevel SEM), and the estimation method [Bayesian estimation vs. frequentist (maximum likelihood) estimation]. It is an interesting topic of future research to compare the outcomes of the two methods.

ALTERNATIVE APPROACHES WITH THREE-LEVEL DATA

Although it seems straightforward to analyse three-level data with the before mentioned approaches as well, I am not aware of any published articles in which measurement invariance with respect to the Level 2 and Level 3 cluster variables is investigated. One option would be to treat the Level 3 clustering as fixed, and impose the measurement invariance restrictions on the two-level models for every school. That is, first measurement invariance across school classes can be investigated using the test for cluster bias (Jak et al., 2013) for each school separately, and next the equality of factor loadings and intercepts can be tested across schools (see Muthén et al., 1997). This approach is not considered very useful, as within each school, the number of school classes will never be large enough to obtain stable estimates and have acceptable power to reject measurement invariance. The BSEM approach can probably be extended to three-level data, by including difference parameters for the intercepts and factor loadings at the school level as well as at the class level. One difference parameter would then reflect how the specific school average differs from the overall average, and another difference parameter would reflect how the specific class deviation from the school average differs from the average class deviation from the school average. The method of Verhagen and Fox could also be extended to three-level data, by estimating school level variance for each measurement parameter.

Although the three-level SEM method is not the only option to investigate measurement bias in three-level data, it is shown in this article that it is at least a relatively simple method to use. At the higher levels of multilevel data, the power of the statistical tests may not be very large, as the number of higher level units is often small. In the current example there were 50 schools at Level 3. From simulation research with two-level data (Jak and Oort, under review), we know that with 50 clusters of size 5, the power to detect large bias is only 50%. Extrapolating this to the three-level situation indicates that that in our example, we did not have high power to detect bias at Level 3. Nevertheless, the illustration can be useful as an example of how the detection of measurement invariance in three-level data may be executed.

ACKNOWLEDGMENT

I am grateful to Johan Schokker from Boom Testuitgevers in Amsterdam for sharing the data with me.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00745/abstract

REFERENCES

Bollen, K. A. (1998). *Structural Equation Models*. New York, NY: John Wiley & Sons, Ltd.

Browne, M. W., and Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociol. Methods Res.* 21, 230–258. doi: 10.1177/0049124192021002005

Butterworth, B. (2005). "Developmental dyscalculia," in *The Handbook of Mathematical Cognition*, ed J. D. Campbell (New York, NY: Psychology Press). 455–469.

De Jong, M. G., Steenkamp, J. B. E., and Fox, J. P. (2007). Relaxing measurement invariance in cross−national consumer research using a hierarchical IRT model. *J. Consum. Res.* 34, 260–278. doi: 10.1086/518532

Devine, A., Soltész, F., Nobes, A., Goswami, U., and Szûcs, D. (2013). Gender differences in developmental dyscalculia depend on diagnostic criteria. *Learn. Instr.* 27, 31–39. doi: 10.1016/j.learninstruc.2013.02.004

Fox, J. P. (2010). *Bayesian Item Response Modeling*: *Theory and applications*. New York, NY: Springer. doi: 10.1007/978-1-4419-0742-4

Fox, J.-P., and Verhagen, A. J. (2010). "Random item effects modeling for cross-national survey data," in *Cross-Cultural Analysis: Methods and Applications*, eds E. Davidov, P. Schmidt, and J. Billiet (London: Routledge Academic), 467–488.

Gross-Tsur, V., Manor, O., and Shalev, R. S. (1996). Developmental dyscalculia: prevalence and demographic features. *Dev. Med. Child Neurol.* 38, 25–33. doi: 10.1111/j.1469-8749.1996.tb15029.x

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: conventional versus new alternatives. *Struct. Equ. Modeling* 6, 1–55. doi: 10.1080/10705519909540118

Jak, S., Oort, F. J., and Dolan, C. V. (2013). A test for cluster bias: detecting violations of measurement invariance across clusters in multilevel data. *Struct. Equ. Modeling* 20, 265–282. doi: 10.1080/10705511.2013.769392

Jak, S., Oort, F. J., and Dolan, C. V. (2014). Measurement bias in multilevel data. *Struct. Equ. Modeling Multidiscip. J.* 21, 31–39. doi: 10.1080/10705511.2014.856694

Kim, E. S., Kwok, O. M., and Yoon, M. (2012). Testing factorial invariance in multilevel data: a Monte Carlo study. *Struct. Equ. Modeling Multidiscip. J.* 19, 250–267. doi: 10.1080/10705511.2012.659623

Kosc, L. (1974). Developmental dyscalculia. *J. Learn. Disabil.* 7, 164–177. doi: 10.1177/002221947400700309

Mehta, P. D., and Neale, M. C. (2005). People are variables too: multilevel structural equations modeling. *Psychol. Methods* 10:259. doi: 10.1037/1082-989X.10.3.259

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Stat.* 13, 127–143.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behav. Res.* 29, 223–236. doi: 10.1207/s15327906mbr2903_2

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Milikowski, M., and Vermeire, S. (2013). *Nederlandse Dyscalculie Screener (NDS). Handleiding en Verantwoording*, Amsterdam: Boom Testuitgevers.

Millsap, R. E., and Everson, H. (1991). Confirmatory measurement model comparison using latent means. *Multivariate Behav. Res.* 26, 479–497. doi: 10.1207/s15327906mbr2603_6

Muthén, B. (1990). *Mean and Covariance Structure Analysis of Hierarchical Data*. Los Angeles, CA: UCLA statistics series, NO. 62.

Muthén, B., and Asparouhov, T. (2013). *New Methods for the Study of Measurement Invariance with Many Groups*. Technical report. Available online at: http://www.statmodel.com

Muthén, B. O., Khoo, S. T., and Gustafsson, J. E. (1997). *Multilevel Latent Variable Modeling in Multiple Populations*. Technical report. Available online at: http://www.statmodel.com

Muthén, L. K., and Muthén, B. O. (1998–2012). *Mplus User's Guide. 7th Edn*. Los Angeles, CA: Muthén and Muthén.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika* 69, 167–190. doi: 10.1007/BF02295939

Stoel, R. D., Garre, F. G., Dolan, C. V., and van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychol. Methods* 11, 439–455. doi: 10.1037/1082-989X.11.4.439

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 2, 4–69. doi: 10.1177/109442810031002

Van De Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthen, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

Verhagen, A. J., and Fox, J.-P. (2012). Bayesian tests of measurement invariance. *Br. J. Math. Stat. Psychol.* 66, 383–401. doi: 10.1111/j.2044-8317.2012.02059.x

Wicherts, J. M., and Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: an illustration using IQ test performance of minorities. *Educ. Meas. Issues Pract.* 29, 39–47. doi: 10.1111/j.1745-3992.2010.00182.x

Widaman, K. F., and Reise, S. P. (1997). "Exploring the measurement invariance of psychological instruments: applications in the substance use domain," in *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*, eds K. J. Bryant, M. Windle, and S. G. West (Washington, DC: American Psychological Association), 281–324. doi: 10.1037/10222-009

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 April 2014; accepted: 26 June 2014; published online: 25 July 2014.
Citation: Jak S (2014) Testing strong factorial invariance using three-level structural equation modeling. Front. Psychol. ***5****:745. doi: 10.3389/fpsyg.2014.00745*
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

METHODS ARTICLE
published: 23 December 2014
doi: 10.3389/fpsyg.2014.01469

Approximate measurement invariance in cross-classified rater-mediated assessments

Ben Kelcey[1]*, Dan McGinn[2] and Heather Hill[2]

[1] College of Education, Criminal Justice and Human Services, University of Cincinnati, Cincinnati, OH, USA
[2] Graduate School of Education, Harvard University, Cambridge, MA, USA

Edited by:
Peter Schmidt, University of Giessen, Germany

Reviewed by:
Daniel Saverio John Costa, University of Sydney, Australia
Pietro Cipresso, Istituto di Ricovero e Cura a Carattere Scientifico Istituto Auxologico Italiano, Italy
Bengt Muthen, Mplus, USA

***Correspondence:**
Ben Kelcey, College of Education, Criminal Justice and Human Services, University of Cincinnati, 2820 Bearcat Way, Cincinnati, 45221 OH, USA
e-mail: ben.kelcey@gmail.com

An important assumption underlying meaningful comparisons of scores in rater-mediated assessments is that measurement is commensurate across raters. When raters differentially apply the standards established by an instrument, scores from different raters are on fundamentally different scales and no longer preserve a common meaning and basis for comparison. In this study, we developed a method to accommodate measurement noninvariance across raters when measurements are cross-classified within two distinct hierarchical units. We conceptualized random item effects cross-classified graded response models and used random discrimination and threshold effects to test, calibrate, and account for measurement noninvariance among raters. By leveraging empirical estimates of rater-specific deviations in the discrimination and threshold parameters, the proposed method allows us to identify noninvariant items and empirically estimate and directly adjust for this noninvariance within a cross-classified framework. Within the context of teaching evaluations, the results of a case study suggested substantial noninvariance across raters and that establishing an approximately invariant scale through random item effects improves model fit and predictive validity.

Keywords: measurement invariance, random item effects, multilevel item response models, teaching, measurement equivalence

The use of rater inferential judgment is a common and persistent feature of assessments designed to measure latent constructs across many different fields of research (e.g., Engelhard, 2002). In these types of assessments, raters typically conduct evaluations by interpreting evidence (e.g., responses, behaviors) using their trained, but subjective, judgments. For this reason, the use of raters to assign scores has been described as an indirect or rater-mediated process because measurements are not directly observed but rather inferred through raters' judgments (Bejar et al., 2006).

An important assumption underlying meaningful comparisons in rater-mediated assessments is that measurement is invariant across raters. Measurement invariance across raters suggests that raters use items similarly so that the relationships between a latent trait and the manifest items with which it is measured do not depend upon which rater conducted an evaluation[1]. When items function differently across raters, ratings no longer preserve a common meaning and basis for comparison across raters because scales are rater-specific. In this way, the extent to which a common scale can be formed across raters depends largely on the extent to which raters share a common basis for assigning scores.

Research has shown that a significant source of construct-irrelevant variation in many rater-mediated assessments arises from differences among raters in how they apply the standards established by an instrument (e.g., Hill et al., 2012). Although findings of rater differences are not surprising, the magnitude and item-specific nature of these differences found by recent reports have demonstrated just how critical of an issue rater variability can be and raises questions about the degree to which scores from different raters are on commensurate scales (Kane and Staiger, 2012). Despite extensive and consistent evidence of rater differences across a broad array of assessments, scores from different raters are routinely treated as if they were exchangeable across raters and are often used to make high-stakes comparative decisions (e.g., Baumgartner and Steenkamp, 2001; Engelhard, 2002; Linacre and Wright, 2002; Eckes, 2009a,b; Schochet and Chiang, 2010; Kane and Staiger, 2012).

In this study, we developed a method to accommodate measurement noninvariance across raters when measurements are nested within raters and (optionally) cross-classified among other distinct hierarchical units (e.g., countries). To do so, we extend cross-classified (multilevel) graded response models to incorporate random item (discrimination and threshold) effects to test, calibrate, and account for measurement noninvariance among raters. By leveraging empirical estimates of rater-specific deviations in the item parameters, the proposed method affords identification of noninvariant items and empirical estimation and direct adjustment for noninvariance within a multilevel or cross-classified framework.

To explore the value of the approach, we applied the proposed method to a case study of repeated classroom measures of teaching quality using three primary questions. First, we investigated the extent to which there was evidence of measurement

[1] We use the term "item" to describe indicators of a latent trait in a broad sense.

noninvariance among raters in cross-classified rater-mediated assessments of teaching. Second, we examined the extent to which allowing item parameters to vary across raters improved the relative and absolute fit of the measurement model as compared to models that assume invariant item parameters. Finally, because a primary criterion for the validity of classroom observations is their efficacy in predicting student achievement gains, we assessed the extent to which allowing item parameters to vary across raters improved the predictive efficacy of observation scores as compared to more conventional approaches.

BACKGROUND

RATER-MEDIATED ASSESSMENTS

Raters have played a critical role in evaluating a wide range of psychological, cognitive, and physical traits. For example, teachers have been used as raters to assess students' medication use and deviant behavior (Conners, 1969; Werry et al., 1975); teachers have been used to rate children's levels of hyperactivity (Gordon, 1979); college instructors have been used to rate students' writing quality (Sudweeks et al., 2004); school principals or trained raters have been used to describe and evaluate teaching through portfolios, instructional diaries, and classroom observations (Brophy, 1986; Kane and Staiger, 2012).

The impetus for the use of rater-mediated assessments stems largely from the position that they often allow for more authentic and relevant assessments, thereby improving support for the validity of an assessment. Despite the flexibility and authenticity offered by rater-mediated assessments, they are often paired with features that, without proper treatment, can undermine their validity and reliability. In particular, a key threat to their validity is the construct-irrelevant variance introduced by the differences among raters in how they award scores (Messick, 1989).

Research across multiple disciplines has demonstrated that such differences manifest in a number of common ways. Perhaps the most commonly cited rater effect is the differences among raters in terms of the severity with which they apply their evaluations. Differences in severity occur when some raters provide ratings that are consistently more severe relative to other raters (Linacre and Wright, 2002). More complex differences of this type can also take root when, for example, rater severity varies across items and/or categories within items. For instance, for a given item some raters may perceive the implied proficiency levels of two adjacent ratings to be further apart than other raters do (Eckes, 2009b). Other common rater effects include a halo effect and a central/extreme tendency effect. Rater halo effects can occur when raters place undue emphasis on a specific competency (Engelhard, 2002). Central/extreme tendencies manifest when raters avoid or use only the extreme categories of a scale (Baumgartner and Steenkamp, 2001).

Together these and other inconsistencies across raters potentially introduce measurement noninvariance because the categories of a scale may no longer have a consistent meaning across raters. Left untreated, rater noninvariance has the potential to unfairly affect outcomes and undermine the reliability and validity of rater-mediated assessments (Messick, 1989).

ANALYSIS OF RATER-MEDIATED ASSESSMENTS

There are a wide variety of approaches to analyzing rater-mediated assessments (e.g., Baumgartner and Steenkamp, 2001; Engelhard, 2002; Patz et al., 2002; Wolfe, 2004; Bejar et al., 2006; De Jong et al., 2007; Lahuis and Avis, 2007; Hill et al., 2012; Carlisle et al., 2013). We focus our discussion on one common treatment of rater-mediated assessments that draws on multilevel measurement models to track rater differences through random effects (e.g., Lahuis and Avis, 2007). We consider two general data structures that are relevant to the proposed model and conceptually outline the application of multilevel measurement models to these data structures.

Hierarchically nested assessments

In studies focused on the comparative evaluation of individuals (e.g., examinees, respondents), assessments are often obtained through the judgmental scoring of participants on targeted indicators (e.g., behaviors, responses) by individual judges. The structure of this design is often considered to have a multilevel organization because participants are hierarchically nested within raters. As previously noted, an important implication of this design is that, to the extent that raters vary in their application of the instrument standards, participants judged by the same rater share construct-irrelevant variation owing to differences among raters. As a result, the nested structure of this design potentially confounds variation in the underlying construct with differences among raters because variation in awarded scores incorporates variation owing to both of these components.

Because the goal of rater-mediated assessments is to assess participants free of rater influence, research has accounted for rater differences by introducing rater effects through, for example, a multilevel item response theory framework (e.g., Lahuis and Avis, 2007). For instance, using an item response model (IRM) where items are incorporated as fixed effects, associations among items are decomposed into a component due to the targeted latent trait and a component designed to capture persistent differences among raters in terms of their relative severity across all items. Given dichotomous items, we might express the probability of receiving a rating of one on item i in for participant t rated by rater r as following a multilevel IRM (where Φ is the normal cumulative distribution function).

$$P(Y_{itr} = 1) = \Phi(a_i\theta_t + a_i\gamma_r - d_i) \quad (1)$$

Here, the probability of obtaining a one on an item is specified as a function of the level of the targeted construct for participant t, θ_t, and the severity of the assigned rater, γ_r, with associated item parameters, a_i as the discrimination parameter and d_i as the threshold parameter. Both latent variables are generally assumed to have a normal distribution and the scale can be set by fixing the distribution of $\theta_t \sim N(0, 1)$.

Cross-classified assessments

Separate from the nesting of participants in raters, rater-mediated assessments frequently introduce, or sustain other design features that further contribute to construct-irrelevant variance. For instance, repeated measures designs are often purposefully

employed in conditions where measurement is known to be unreliable or sensitive to context (Hill et al., 2012). Similarly, many measurement designs operate within larger multilevel structures. For example, participants may be nested within schools or nested within countries (Steenkamp and Baumgartner, 1998; Fox, 2010).

A common result of these design features is that they introduce a cross-classified dependence structure in the data because each participant or observation is simultaneously nested within a rater and a second distinct non-hierarchical unit (Baayen et al., 2008). For example, under a repeated measures design, each participant is observed across multiple observations and each observation is rated by a different rater. Observations are thus nested within or cross-classified among participants and raters.

Under the repeated measures design, research has found evidence that scores among items within the same observation are likely to display excess variance arising from rater differences and idiosyncratic features of an observation (e.g., participant had a bad day). Because such excess variance is specific to an observation and rater and does not generalize beyond a sampled observation and rater, research has accounted for these effects by introducing observation- and rater-specific random effects (e.g., Carlisle et al., 2013). The introduction of random effects for each mode of the distinct hierarchies gives rise to a cross-classified (multilevel) IRM. Variation in the targeted latent trait is now decomposed into three components: a targeted participant component which persists across observations, an observation-specific component, and a rater component. Extending the multilevel IRM in Equation (1), we can now express the probability of obtaining a particular rating as

$$P(Y_{iotr} = 1) = \Phi(a_i\theta_t + a_i\alpha_{ot} + a_i\gamma_r - d_i) \qquad (2)$$

Equation (2) follows the aforementioned notation but now expands to accommodate (a) repeated measurements such that Y_{iotr} is the score on item i in observation o for participant t rated by rater r and (b) observation-specific deviations for observation o in participant t (α_{ot}).

APPROACHES TO MEASUREMENT INVARIANCE

To assess and substantiate invariance in these applications or correct for noninvariance, there have been three typical approaches: full, partial, and approximate invariance. Below we briefly outline their structure, application, and limitations as they may apply to rater-mediated assessments.

Full invariance

The conventional approach to assessing/establishing invariance across subgroups is through multiple group analyses. For instance, continuing with the aforementioned notation from the repeated measures cross-classified model (2), full invariance across raters supports

$$P(Y_{iotr} = 1|R, \theta) = P(Y_{iotr} = 1|\theta) \qquad (3)$$

(Mellenbergh, 1989). Put differently, for participants with the same level of the latent trait, the probabilities of a particular score on an item should not depend on which rater rated an observation (Millsap and Everson, 1993).

Partial invariance

Measurement becomes noninvariant when the relationships between a latent trait and items depend on which group an observation belongs to [e.g., the equality in Equation (3) no longer holds]. When there is evidence of measurement noninvariance, a common alternative approach is to adjust for noninvariant items using a partial measurement invariance approach (Steenkamp and Baumgartner, 1998). With partial measurement invariance, multiple group (e.g., rater-specific) measurement models are estimated and linked to form a common scale (across groups) by capitalizing on items that are invariant across all groups (i.e., anchor items). Despite the potential of the partial measurement invariance approach, literature has highlighted several important limitations (e.g., Holland and Wainer, 1993; Vandenberg, 2002; Steinmetz, 2013). Perhaps most germane to multilevel and cross-classified rater-mediated assessments is that empirical application of a partial invariance approach requires invariant items across all groups in order to bridge groups-specific scales. Lacking invariant items to anchor the scale across raters, multigroup partial invariance approaches are poorly suited to establish a common scale across groups (e.g., Holland and Wainer, 1993). Furthermore, even if two invariant items existed, estimating and testing for such invariance with a multigroup model would conceptually require estimating a separate measurement model for each rater. Given a large number of raters, stable estimation of item parameters would likely require large sample sizes and be computationally demanding because of the number of estimated parameters.

Approximate measurement invariance

When full or partial measurement invariance is intractable, a more flexible approach recently developed is to accommodate measurement noninvariance through hierarchically defined random item effects (Fox, 2010; Rijmen and Jeon, 2013). The prototypical application involves cross-national comparisons of latent traits with respondents nested within countries (Fox, 2010). To facilitate cross-national comparisons, measurement invariance requires items to function similarly in each country. When items are not invariant across countries, the approximate measurement invariance approach uses random item effects to model the extent to which item parameters vary across countries. This approach establishes an international measurement scale across countries using the mean of item parameters across all countries. Country-specific noninvariance in item parameters is then conceptualized as deviations from the international item parameters and captured through country-specific random item effects.

There are two primary practical advantages to this framework. First, in theory, a common scale can be established and cross-group comparisons can be made even when no items are strictly invariant across countries (Fox, 2010). Second, because the framework draws on random instead of fixed item effects, it presents a much more parsimonious representation of the differences among groups in terms of estimated model parameters. Investigations that include many groups are more feasible because

the number of estimated parameters does not increase rapidly with the number of groups.

A nascent but growing body of research has demonstrated the potential of this approach (De Boeck, 2008; Muthén and Asparouhov, 2013). Simulation studies have shown that the multilevel random item effects framework recovers both overall and group-specific item parameters well in a variety of settings (Fox and Verhagen, 2010). Similarly, simulations assessing the comparative performance of invariance approaches have suggested that the approximate measurement invariance approach outperforms full and partial invariance approaches when there are many small differences in item parameters (Van de Schoot et al., 2013). Substantive applications have also emphasized the value of multilevel random item effects methods in accounting for response heterogeneity across groups (De Jong et al., 2008; Fox and Verhagen, 2010).

MODEL FORMULATION

When an IRM, such as those noted above, fit the data, we can separate estimates of the targeted latent trait from the distributional properties of items such that estimates generalize beyond the sampled observations and raters (Linacre, 1989). The critical assumption that allows for the separation of the latent trait from item characteristics is that measurement is invariant across subgroups of a population (Van de Schoot et al., 2013). Given a multilevel or cross-classified data structure, the conditions underlying the validity of this separation require invariance across each facet (e.g., participants, raters, observations).

More conceptually, construct-irrelevant variation can be split into two principal sources—latent trait side variation and item side variation. Latent trait side construct-irrelevant variation arises when the actual latent trait varies across design facets such as raters and/or observations. In contrast, item side variation arises when the underlying relationships between items and a latent trait vary across, for example, raters.

Under this division of construct-irrelevant variation, the aforementioned measurement models (Equations 1, 2) solely address latent trait variation across facets because they (only) decompose the variation in a latent trait into components uniquely attributable to each facet and do not address how item parameters vary across facets. Put differently, the latent trait side random effects models presented above account for the extent to which the latent trait of a participant is deflected by, for example, the relative severity of a rater and/or the atypical nature of an observation. In this way, latent trait side random effects models accommodate threshold differences among raters and observations only if these differences manifest consistently and uniformly for all items. If rather threshold differences among raters/observations vary across items or if discrimination parameters differ, latent trait side random effects models will not be sufficient to separate the latent trait from item characteristics because measurement is not invariant across facets.

Rather, in the presence of item side variance, separation of the latent trait from item characteristics would require direct treatment of measurement noninvariance. Applied to cross-classified rater-mediated assessments, conventional approaches, such as the partial invariance approach, are however particularly challenging because studies tend to draw on large number of raters and only a small number of items per latent trait. To relax assumptions of measurement invariance across raters, we developed a random item effects cross-classified (multilevel) graded response model. Our specification first drew on a graded response model parameterization such that observed item scores were treated as fallible ordinal ratings stemming from a targeted latent trait. Second, because many rater-mediated assessments operate within cross-classified (multilevel) designs, we leveraged a cross-classified (multilevel) graded response model to introduce random effects for distinct hierarchical units (e.g., raters). Third, we accommodated noninvariance across raters by permitting item discrimination and threshold parameters to vary across raters (and potentially another hierarchical unit) using random item effects (Fox, 2010). Under a repeated measures design, we express our model as

$$P(Y_{iotr} = k) = \Phi(a_i\theta_t + a_{ir}\alpha_{ot} + a_{ir}\gamma_r - d_{ir}^{k-1}) - \Phi(a_i\theta_t + a_{ir}\alpha_{ot} + a_{ir}\gamma_r - d_{ir}^{k}) \quad (4)$$

Here Y_{iotr} is the ordinal score for item i in observation o for participant t rated by rater r, a_i represents the average discrimination parameter for item i across all raters, θ_t represents a participant's persistent level of the targeted latent trait (i.e., across all observations), a_{ir} is item i's discrimination parameter under rater r, α_{ot} is the latent trait deviation specific to observation o for participant t, and γ_r is the deviation capturing consistent differences among raters in terms of their relative severity across all items. Let K represent the number of categories items are graded on with k as a specific category and let $d_{ir}^{(1)}$,..., $d_{ir}^{(K-1)}$be a set of K–1 ordered item thresholds. That is, γ subsumes threshold differences among raters that are consistent across items, whereas d captures threshold differences among raters that are item-specific. To set the scale, let $\theta \sim N(0, \sigma_t^2)$, $\alpha \sim N(0, 1)$, $\gamma \sim N(0, \sigma_r^2)$, $a_{ir} \sim N(a_i, \sigma_{a,i}^2)$, and $d_{ir}^k \sim N(d_i^k, \sigma_{d,i}^2)$.

In this particular specification, we used an independent random item effects structure and restricted item parameters to vary across only a single level two unit (raters). However, the model could be further extended to consider covariance among random item effects parameters and/or to allow item parameters to vary across both level two units (e.g., raters and participants). Similarly, we applied the mean item parameters across raters as the inter-rater item parameters and use these to construct an inter-rater scale. However, there are many reasonable and potentially more appropriate alternatives.

For instance, one alternative specification estimates the discrimination parameter applied to a participant's persistent level of the targeted latent trait (θ_t) separate from the observation level discrimination parameter (a_{ir}).

$$P(Y_{iotr} = k) = \Phi(a_i^{(t)}\theta_t + a_{ir}^{(o)}\alpha_{ot} + a_{ir}\gamma_r - d_{ir}^{k-1}) - \Phi(a_i^{(t)}\theta_t + a_{ir}^{(o)}\alpha_{ot} + a_{ir}\gamma_r - d_{ir}^{k}) \quad (5)$$

Here we now use $a_{ir}^{(o)}$ as the observation level discrimination parameters (where $a_{ir}^{(o)} \sim N(a_i^{(o)}, \sigma_{a,i}^2)$) and introduce $a_i^{(t)}$ as the participant level discrimination parameters which are nonrandom and unconnected to the observation level discrimination

parameters. Under this specification, the scale of θ_t can be set by fixing its distribution to $\theta \sim N(0, 1)$.

The proposed model can also be adapted to accommodate other cross-classified or multilevel structures. For example, as noted earlier, many measurement designs operate within larger multilevel structures. Consider for example a design in which participants are cross-classified among raters and schools in which we track measurement noninvariance across raters. Under this design, the targeted latent trait of a participant now operates at lowest level of the hierarchy. With some slight changes in notation we can modify Equation (4) so that

$$P(Y_{itsr} = k) = \Phi(a_i\theta_s + a_{ir}\alpha_{ts} + a_{ir}\gamma_r - d_{ir}^{k-1}) - \Phi(a_i\theta_s + a_{ir}\alpha_{ts} + a_{ir}\gamma_r - d_{ir}^{k}) \quad (6)$$

Here Y_{itsr} is the ordinal score for item i of participant t in school s rated by rater r, a_i represents the average discrimination parameter for item i across all raters, θ_s represents the school effect or school-specific deviation in the latent trait, a_{ir} is item i's discrimination parameter under rater r, α_{ts} is participant t's level of the targeted latent trait, and γ_r is the deviation specific to rater severity. Remaining notation and constraints are unchanged.

Our formulation of approximate measurement invariance models for rater-mediated assessments within a cross-classified (multilevel) structure is an extension of the multilevel IRM with random item effects (Fox, 2007). The proposed method first conceptualizes rater-mediated assessments and differential item functioning across raters within a multilevel random item effects framework. In turn, the method extends strictly hierarchical structures to accommodate cross-classified data structures where level one units (e.g., observations) are simultaneously nested within two independent level two units (e.g., raters and participants). Subsequently, we used this cross-classified framework to introduce hierarchically defined latent variables for both the targeted construct and the items to capture their respective variability across distinct level two units.

As noted earlier, construct-irrelevant variation can be conceptually split into two principal sources—latent trait and item side variation. Latent trait random effects (e.g., Equations 1, 2) serve to decompose the variation in a latent trait across facets. In contrast, item side random effects serve to capture the extent to which items function differently across hierarchical units. By simultaneously introducing latent trait and item side random effects, we permit a latent trait to vary across hierarchical units and items to function differently across those hierarchical units. When the proposed model fits the data, decomposing the latent trait and adjusting for differential item functioning across raters through random effects can establish an inter-rater scale such that the latent trait is separable from construct-irrelevant variation. In this way, estimates of a targeted latent trait from models that accommodate both latent trait and item side variation are more likely to generalize beyond the sampled observations and raters.

The key addition in the approach is the introduction of item side random effects across raters within a cross-classified framework. Random item effects are intended to not only identify noninvariance but also to track it through empirical estimates of the differences among raters. Under a Bayes approach, empirical estimates of rater-specific differences in item parameters are obtained using a mix of the inter-rater item parameters, which are based on all observations, and rater-specific item parameters, which are based on the particular observations a rater has rated. Rater-specific differences in item parameters are estimated using a shrinkage estimator where the amount of shrinkage toward the inter-rater estimates is a function of how precisely we can identify raters' differences from the mean. In this way, random item effects allow us to borrow strength from the larger pool of raters to improve estimates for individual raters, especially those for which we have little information. The shrinkage of rater-specific item parameters toward inter-rater parameters has been shown to reduce the mean-squared error of rater-specific estimates and is widely used elsewhere (Lindley and Smith, 1972; Raudenbush and Bryk, 2002; Fox, 2010).

In situating the proposed repeated measures model (Equation 4) among more conventional models, a single level IRM assumes that associations among items derive solely from a targeted latent trait. A multilevel IRM with observations nested within participants (ignoring raters) suggests that associations among items derive from a persistent component of a targeted latent trait and observation-specific deviations. Use of a cross-classified IRM with observations cross-classified among raters and participants suggests associations among items are a function of a persistent component of a targeted latent trait, observation-specific deviations, and deflections due to consistent differences in severity among raters. In these latent trait side (only) random effects models, item parameters are assumed to remain equal across raters. If we further introduce random item effects into the cross-classified model (Equation 4), we relax this assumption of equality of item parameters across raters and allow the discrimination and threshold parameters to vary.

ESTIMATION

The cross-classified structure of this model combined with the potential for a large number of latent variables renders maximum likelihood estimation computationally challenging with even a few items because it would require high dimensional numerical integration. A more practical option in this context is Bayesian methods (Gelman et al., 2004; Fox, 2007; Asparouhov and Muthén, 2012). Albert and Chib (1993) described a Gibbs sampler for a graded response model by using normally distributed latent item responses, Z_{iotr}. Under this formulation, an observed ordinal response, Y_{iotr}, is used as an item of a normally distributed latent item response, Z_{iotr}, which is placed into a response category defined by threshold parameters d_{ir}^{k} such that Z_{iotr} is defined as

$$Z_{iotr}|Y_{iotr} = k, \theta_t, \alpha_{ot}, \gamma_r, d_{ir}^{k}, d_{ir}^{k-1}, a_{ir}, a_i \sim N(a_i\theta_t + a_{ir}\alpha_{ot} + a_{ir}\gamma_r, 1)I(d_{ir}^{k-1} < Z_{iotr} \leq d_{ir}^{k}) \quad (7)$$

This framework and its variations have been extended to incorporate multilevel structures and can be implemented in, for

example, Mplus (De Jong et al., 2007; Asparouhov and Muthén, 2010a,b; Fox, 2010; Muthén and Muthén, 1998–2012).

TESTING FOR NONINVARIANCE

Having introduced random item effects to accommodate measurement noninvariance across raters, a relevant question is how we might test for evidence of (non)invariance. If measurement invariance holds, the variance of the random item effects across raters should be zero (e.g., $\sigma^2_{a,i} = 0$). That is, if the variance of the random item effects is zero, item parameters are consistent across raters and measurement is invariant. However, departures from zero for specific items suggest that measurement is noninvariant across raters because the relationship between an item and the latent trait is not consistent across raters.

To examine evidence for measurement invariance and assess relative model fit, we can employ Bayesian tests of measurement invariance (Verhagen and Fox, 2013). These tests evaluate the variance components of the random item effects by using the Bayes factor to compare the ratio of the marginal likelihood of the null model (invariance) with the marginal likelihood alternative (noninvariance). Within the context of random item effects models, Bayesian tests of measurement invariance can be used to test invariance for each item parameter simultaneously by comparing models estimated with a diffuse prior against those using an informative prior concentrated at zero (e.g., inverse gamma distribution with a small scale parameter). Such comparisons potentially identify differential item functioning and directly assess the extent to which the fit of a model with fixed item parameters is improved upon by allowing item parameters to vary. Additional tests of, for example, factor variance invariance can also be investigated (e.g., Steenkamp and Baumgartner, 1998).

APPLICATION

To probe the potential value and utility of the proposed methods, we applied our proposed model to a study of teaching quality using repeated classroom observations of mathematics teaching. As noted earlier, we investigated three questions focused on (a) evidence of noninvariance, (b) improvements in relative and absolute fit, and (c) improved predictive validity. Although we use this application as an initial case study of the proposed method, we are cautious to note that the correct underlying model is unknown because it is an empirical investigation. For this reason, the extent to which differences among approaches represent true gains or the extent to which these gains might be generalizable is unknown and needs to be studied further.

DATA DESCRIPTION

In assessments of teaching quality, classroom observations of teaching are generally carried out by having trained raters evaluate teachers across multiple observations using a fixed set of items. Teaching evaluation instruments typically focus raters' attention on behaviors that exemplify an implicit theory of effective teaching. For each item, the guiding rubric that accompanies each instrument typically provides specific examples and descriptive anchors for each category of a scale and raters typically provide ordinal assessments for each item in each observation.

Like other types of rater-mediated assessments, a significant source of construct-irrelevant variation in classroom observations is differences among raters in their judgments (Kane and Staiger, 2012). The issue of rater differences can be especially pronounced in modern classroom observation systems because, unlike their historical counterparts, modern systems go beyond simple low inference checklists and rely more on inferential judgments. Recent investigations have demonstrated that even with extensive rater training, substantial differences among raters persist (Bell et al., 2012; Hill et al., 2012; Kane and Staiger, 2012).

Our data on teaching quality came from the National Center for Teacher Effectiveness study, which focused on identifying teacher characteristics and teaching practices that correlate with teacher effects as measured through student test score outcomes. Data for this analysis focus on classroom observations across two academic years of 150 fourth- and fifth-grade mathematics teachers and their students situated within across four large urban school districts in the Eastern United States. Each observation lasted about an hour and teachers were observed over three different occasions across an academic year. For each of these occasions, teachers were rated using the Mathematical Quality of Instruction (MQI) classroom observation system (Hill et al., 2008).

Teacher quality measure

The MQI observation system is a subject-specific observation instrument that was designed to provide a balanced view of mathematics instruction (Hill et al., 2008). In the current investigation, we focused our analyses on a general teaching quality domain which was captured using four ordinal items. The first item measured the extent to which the observed classroom work was consistently and directly connected to mathematics content (CWCM). The second item, richness of the mathematics instruction (RICH), captured the depth of the mathematics offered to students (Hill et al., 2008). The third item, Working With Students (WWS), captured the quality with which teachers understand and respond to students' mathematically substantive productions. The final item measured student participation in meaning-making and reasoning (SPMMR). This item captured students' involvement in cognitively demanding tasks and the extent to which students participated in and contributed to meaning-making and reasoning.

For each observation, raters independently evaluated teachers' instruction along each of the items by grading them on an ordinal scale ranging from a low of one to a high of three according to the descriptive anchors provided by the MQI rubric. The only exception was the CWCM item which was dichotomous. As a result, evaluations for each observation consisted of ordinal scores on a fixed set of items with each observation cross-classified by two hierarchical grouping structures—teachers and raters.

Each of the 39 raters in this study completed an online MQI training program (approximately 16 h) and then passed a subsequent certification exam. Raters also completed weekly calibration exercises where their scores were compared to master scores on clips of instruction. These scores were discussed in weekly webinars with master raters to help prevent rater drift. Raters who demonstrated problematic scores or rationales were remediated by master raters.

Student achievement measure

To measure student achievement, we used a researcher developed test administered to students in all four districts during the fall and spring semesters of the 2010–11 and 2011–12 school years. Items on this low-stakes mathematics assessment were designed to align with fourth and fifth grade Common Core mathematics standards, and covered topics such as numbers and operations, algebra, and geometry and measurement. Reliability of the test ranged from 0.82 to 0.89, depending on the form (Hickman et al., 2012).

To measure the average student achievement gains associated with each teacher in our sample, we estimated the following hierarchical linear model.

$$a_{j,t,f} = A_{j,t,f-1}\pi + X_{j,f}\beta + \xi + \mu_t + \zeta_{t,f} + \varepsilon_{j,t,f} \qquad (8)$$

The outcome variable, $a_{j,t,f}$, represents the performance on the mathematics assessment of student j taught by teacher t, at time f. The model conditioned on a vector of prior achievement measures, $A_{j,t,f-1}$, which includes a cubic polynomial term for prior achievement on the same assessment[2], a standardized English assessment, and their classroom aggregates; time varying demographic indicators, $X_{j,f}$, for student j at time f (which include race, gender, subsidized-lunch eligibility, English language learner status, and special education status; and indicators for district, grade, and year of the assessment, ξ); and residual effects for the teacher (μ_t), time($\zeta_{t,f}$), and student ($\varepsilon_{j,t,f}$). To estimate the underlying teacher effect or "value-added" score, we used the empirical Bayes residual for each teacher.

METHOD

We applied the previously described random item effects cross-classified graded response model (Equation 4). We estimated the models in Mplus using the default diffuse prior distributions (see Appendix). Prior distributions for the discrimination parameters were normal with mean zero and variance five; for the thresholds the prior distributions were normal with mean zero and infinite variance, and for the variance parameters the prior distributions were log uniform bounded by negative and positive infinity. Subsequent inferences were conducted on the posterior medians and standard deviations. For each model, we ran two chains using a burn-in of 25,000 MCMC iterations and up to 100,000 post-burn-in iterations with convergence determined by the default potential scale reduction criteria implemented in Mplus and Gelman-Rubin diagnostics (Gelman and Rubin, 1992).

To assess evidence of measurement noninvariance, we first examined the variances of the item effects and their posterior distributions. To further appraise evidence for measurement invariance and assess relative fit, we employed the aforementioned Bayesian tests of measurement invariance for the null hypothesis that the variance of each item parameter was zero. To do so, we re-estimated the random item effects models using an inverse gamma (informative) prior with a shape parameter value of one and a scale parameter value of 0.005. We then explored the absolute fit using simple posterior predictive checks (Gelman et al., 2004). Finally, we evaluated the predictive capacity of the models by correlating teaching quality with value-added scores. Throughout the analyses we compared the results of the random item effects cross-classified graded response model with the results of alternative models which assume measurement invariance to assess the potential differences across models.

RESULTS

Table 1 presents the posterior item parameter estimates (on a probit scale) from a single level, a multilevel (occasions nested within teachers), a cross-classified (occasions nested within teachers and raters), and a random item effects cross-classified graded response models (Equation 4). For each model without random item effects, we present the item parameters and their uncertainty as captured by the posterior standard deviation. For the model which incorporates random item effects, we include the inter-rater item parameters and the uncertainty of those means using the posterior standard deviation. In addition, we summarize the variability of the item parameters across raters and 95% posterior intervals because the distributions of variance estimates are frequently skewed.

The results of the random item effects model suggested that the item discrimination and threshold parameters varied across raters and thus were noninvariant (**Table 1**). Based on their posterior distributions, 95% posterior intervals suggested that the variance of their discrimination and threshold parameters was significantly different than zero. When the magnitude of item side variation across raters for each item is placed alongside the variance of the latent trait attributable to raters, the results suggested item side variation for each item was about half as large. That is, the variance in the latent trait across raters was about 0.26 (see last row of **Table 1**) whereas the average variance of item parameters among raters across all items was 0.13 (average of item variances in **Table 1**).

To put this into context, consider the Richness item. The estimated variance implies that although the item discrimination parameter was on average about 1.05 across all raters, the discrimination parameter for this item varied depending on who rated an observation (**Table 1**). For a rater who is two standard deviations above average, the estimated discrimination parameter could be as high 1.67 (using double the square root of the "Item Variance Across Raters" column in **Table 1**). In contrast, a rater who is two standard deviations below average, the estimated discrimination parameter for the same item could be as low as 0.43.

To illustrate the implications of this noninvariance, **Figure 1** describes the item characteristic curves across raters for the richness item for the first threshold. In this figure, the dark curve represents the inter-rater item characteristic curve which is the average across all raters. In contrast, the gray curves describe the item characteristic curves for raters who are approximately one or two standard deviations above or below the average discrimination and threshold estimates for this item. Evident from this figure, which rater rates an observation has important implications for the scale of ratings and the extent to which teachers are placed on a similar scale.

[2]One district did not take the study distributed assessment in the fall semester of school year 2010–11 (pretest), so we used student performance on the state standardized math exam in its place for this prior achievement control.

Table 1 | Discrimination and threshold parameters.

Parameter	Single		Multilevel		Cross-classified		Random item effects cross-classified				
	Est	*SD*	Est	*SD*	Est	*SD*	Est	*SD*	Item variance across raters	Low	High
DISCRIMINATION (a_i)											
RICH	1.14	0.04	1.08	0.04	0.99	0.05	1.05	0.07	0.10	0.05	0.20
WWS	1.39	0.07	1.18	0.06	1.15	0.05	1.46	0.11	0.19	0.08	0.44
CWCM	0.79	0.06	0.78	0.07	0.76	0.06	0.74	0.09	0.08	0.02	0.21
SPMMR	1.33	0.06	1.23	0.05	1.17	0.06	1.16	0.07	0.11	0.05	0.23
THRESHOLD (d_i)											
RICH(1)	0.61	0.03	0.72	0.07	0.74	0.12	0.56	0.12	0.12	0.06	0.24
RICH(2)	2.57	0.06	2.88	0.08	2.93	0.14	2.71	0.13			
WWS(1)	0.53	0.04	0.57	0.04	0.64	0.13	0.48	0.15	0.07	0.01	0.24
WWS(2)	2.75	0.10	2.80	0.07	2.94	0.14	3.12	0.20			
CWCM(1)	−1.98	0.06	−2.24	0.12	−2.25	0.15	−2.39	0.15	0.09	0.02	0.25
SPMMR(1)	0.83	0.04	0.94	0.08	1.03	0.16	0.83	0.13	0.25	0.13	0.49
SPMMR(2)	2.78	0.09	3.06	0.12	3.24	0.19	2.97	0.15			
LATENT TRAIT VARIANCE											
Observations	1.00	—	1.00	—	1.00	—	1.00	—			
Teachers	—	—	0.34	0.05	0.40	0.06	0.32	0.06			
Raters	—	—	—	—	0.28	0.09	0.26	0.09			

Est, estimate; SD, standard deviation; Item Variance Across Raters, the item-specific random effect variance across raters ($\sigma^2_{a,i}, \sigma^2_{d,i}$); Low and High, the lower and upper bounds of the 95% posterior interval respectively.

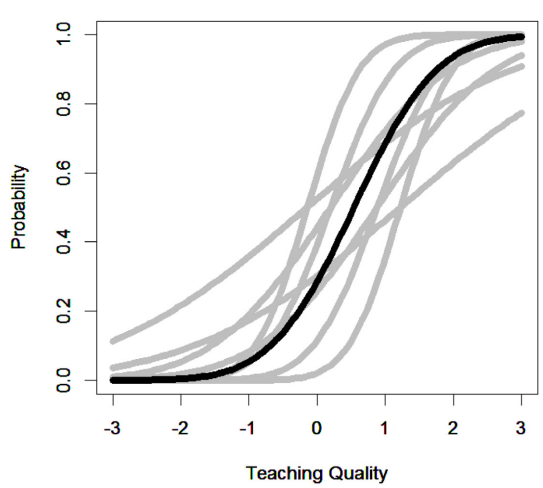


FIGURE 1 | Item characteristic curve for a single item across different raters.

To formally test measurement invariance across raters for each item and to assess relative fit, we re-estimated the random item effects model using an inverse gamma prior distribution of $IG(1, 0.005)$ for the variance of each item parameter to test the null hypotheses that each of the variances was less than 0.001, 0.01, or 0.1. Using a common cutoff of about three for Bayes factor, the results for each threshold and discrimination parameter uniformly indicated that the variance of the random effects was different than zero (Jeffreys, 1961). In **Table 2**, we present the estimated variance along with the bounds of its 95% credible intervals and the Bayes factors for each item parameter under the hypotheses that the respective variance is less than 0.001, 0.01, or 0.1.

We further examined the fit of the models using posterior predictive checks for items. Overall, we found little difference across models. **Table 3** contrasts the observed probability for each category by each item with the model based predicted probability for each model. In each case, the model largely recovers the observed probabilities. The multilevel model slightly misestimated probabilities for the RICH and CWCM items, the cross-classified model without random item effects slightly misestimated the RICH and SPMMR items, and the random item effects cross-classified model slightly misestimated the CWCM item.

To further contrast the methods, we examined the correspondence of their teaching quality estimates. We first examined the correlation among scores from alternative methods. Results indicated that estimates from alternative methods were correlated with the proposed method between 0.89 and 0.93 (**Table 4**). Next, we considered the discrepancy among implied teacher classifications. Current and forthcoming policy often requires that teachers be stratified into about four categories (e.g., Hansen et al., 2013). For each set of scores we classified teachers into quartiles and identified the percentage of discrepant classifications. Results indicated that discrepancy rates between the proposed method and the alternative methods were relatively high and ranged from 23 to 37% (**Table 5**). Put differently, based on a sample size of 150 teachers, approximately 35 to 56 would be classified differently across methods.

As noted earlier, a primary benchmark for the validity of classroom observations is their efficacy in predicting student achievement gains. To examine our final research question, we investigated the extent to which allowing item parameters to vary across raters improved the predictive validity of the teaching observation scores as compared to more conventional approaches. To get a sense of the extent to which improvements in predictive validity were attributable specifically to random item effects across raters, we examined correlations for models that sequentially introduced key features. **Table 6** displays the correlations between teachers' value-added scores and their teaching quality estimates from the single level, multilevel, cross-classified, and random item effects cross-classified models.

The results suggested gains as models increasingly took into account integral features of classroom observation data. Using simple averages, the correlation between observation and value-added scores was 0.11. By applying item response theory and acknowledging the ordinal nature of the scale, this correlation increased by about 10%. By introducing random observation effects through a multilevel model to account for the dependence of items within an observation, the correlation increased an additional 30%. In contrast, further introducing a random effect for raters through a cross-classified model (but restricting item parameters to be invariant across raters), decreased the correlation by about 10%. However, once we allowed for random item effects, the cross-classified model again increased the correspondence between observation and value-added scores. Moreover, although 95% intervals for the correlation between observation and value-added scores included zero across models without random item effects, the 95% interval for the correlation excluded zero for the model with random item effects.

DISCUSSION

Although strict measurement invariance across raters is optimal, the reality is that it will rarely hold in rater-mediated assessments. Developing measurement models that are more tightly attuned to the types of measurement errors present in rater-mediated assessments is likely to improve the validity and comparability of scores across raters and other sources of construct-irrelevant variation. The proposed method relaxes assumptions of measurement invariance in cross-classified (multilevel) rater-mediated assessments by introducing random item effects to test for non-invariance and empirically construct an inter-rater scale. More conceptually, the approach helps to identify the "ruler" each rater uses to conduct his/her assessments, construct an inter-rater scale,

Table 2 | Test of measurement invariance for item parameters.

Parameter	Variance	Low	High	*BF* < 0.001	*BF* < 0.01	*BF* < 0.1
THRESHOLD						
RICH	0.12	0.06	0.24	0.000	0.000	0.576
WWS	0.07	0.01	0.24	0.623	0.668	1.009
CWCM	0.09	0.02	0.25	0.053	0.276	0.884
SPMMR	0.25	0.13	0.49	0.000	0.000	0.009
DISCRIMINATION						
RICH	0.10	0.05	0.20	0.000	0.000	0.858
WWS	0.19	0.08	0.44	0.000	0.000	0.353
CWCM	0.08	0.02	0.21	0.172	0.256	0.986
SPMMR	0.11	0.05	0.23	0.000	0.001	0.783

BF, Bayes factor for each item parameter under the hypotheses that the respective variance is less than 0.001, 0.01, or 0.1; Low and High, the lower and upper bounds of the 95% posterior interval respectively.

Table 4 | Correlation among observation scores from different methods.

Method	RIE-CC	CC	ML	Single	Averages
RIE-CC	1.00	0.93	0.91	0.90	0.89
CC	0.93	1.00	0.92	0.91	0.92
ML	0.91	0.92	1.00	0.96	0.95
Single	0.90	0.91	0.96	1.00	0.99
Averages	0.89	0.92	0.95	0.99	1.00

RIE-CC, random item effects cross-classified graded response model; CC, cross-classified graded response model; ML, multilevel graded response model; Single, single level graded response model

Table 3 | Posterior predictive checks for item fit (95% posterior intervals).

Item category	Observed	Single level		Multilevel		Without random item effects		With random item effects	
		Low	High	Low	High	Low	High	Low	High
RICH0	0.656	0.649	0.663	0.648	0.692	0.637	0.709	0.618	0.69
RICH1+	0.344	0.337	0.351	0.308	0.352	0.291	0.363	0.31	0.382
RICH2	0.044	0.042	0.047	0.031	0.043	0.027	0.042	0.034	0.052
WWS0	0.622	0.613	0.629	0.604	0.65	0.597	0.676	0.573	0.656
WWS1+	0.378	0.371	0.387	0.35	0.396	0.324	0.403	0.344	0.427
WWS2	0.053	0.05	0.057	0.043	0.058	0.038	0.059	0.045	0.071
CWCM0	0.060	0.058	0.062	0.042	0.052	0.041	0.061	0.035	0.05
CWCM1	0.940	0.938	0.942	0.948	0.958	0.939	0.959	0.95	0.965
SPMMR0	0.691	0.683	0.698	0.678	0.723	0.676	0.749	0.668	0.738
SPMMR1+	0.309	0.302	0.317	0.277	0.322	0.251	0.324	0.262	0.332
SPMMR2	0.047	0.044	0.05	0.034	0.048	0.027	0.044	0.03	0.047

Table 5 | Discrepant classification rates among methods.

Method	RIE-CC	CC	ML	Single	Averages
RIE-CC	0.00	0.23	0.32	0.37	0.33
CC	0.23	0.00	0.26	0.30	0.32
ML	0.32	0.26	0.00	0.24	0.23
Single	0.37	0.30	0.24	0.00	0.09
Averages	0.33	0.32	0.23	0.09	0.00

RIE-CC, random item effects cross-classified graded response model; CC, cross-classified graded response model; ML, multilevel graded response model; Single, single level graded response model.

Table 6 | Correlation between observations scores and value-added scores.

	Estimate	Low	High
Averages	0.11	−0.05	0.27
Single	0.12	−0.04	0.28
Multilevel	0.15	−0.01	0.31
CC	0.14	−0.02	0.30
RIE-CC	0.17*	0.01	0.33

**Interval excludes zero.*

RIE-CC, random item effects cross-classified graded response model; CC, cross-classified graded response model; ML, multilevel graded response model; Single, single level graded response model.

and make adjustments to observed scores in order to place them on this inter-rater scale.

Evidence from the case study on teaching quality suggested the promise of random item effect models in addressing noninvariance in rater-mediated assessments. The results indicated that measurement was noninvariant across raters for each item and suggested that direct adjustments for this noninvariance through random item effects improved model fit and the predictive validity of the teaching quality. These results are consistent with prior literature in that they suggest that ignoring measurement noninvariance can obscure both the psychometric properties of a scale and the underlying relationships among variables.

As noted previously, the results presented in this study are only based on a single case study and do not necessarily imply these findings will generalize. However, although the authority of the proposed model over alternative models is unclear in our empirical application, the more flexible assumptions of the proposed model with regard to measurement noninvariance would seem to lend greater credence to its results. Nevertheless, the circumstances under which the proposed method outperforms alternative methods need to be systematically studied in greater detail to understand the extent to which findings are robust to key assumptions.

In this regard, we highlight four areas that warrant further study. First, the flexibility of the proposed framework suggests many different alternative forms and we have presented just a few limited forms. For instance, we chose to define inter-rater parameters as the average of item parameters and apply those values to the teacher level construct. However, there are many reasonable alternatives including not linking parameters at hierarchical levels to those at the lower level at all and independently estimating them. Future research will need to investigate alternatives, develop tests for comparing the fits of non-nested models, and examine the extent to which results are robust to these choices.

Second, in our application we assumed random item effects were independently normally distributed. For our case study, *post-hoc* analyses examining the tenability of the normality assumption for each item parameter using the Shapiro–Wilks test of normality were conducted. Each test suggested that we could not reject the null hypothesis that the random item effects came from a normal distribution. However, this assumption may be untenable if, for example, items are invariant across most raters but demonstrate substantial invariance for a handful of raters. In this case fixed multiple group approaches are potentially more appropriate. Similarly, its reasonable to suspect that random item effects may not be independent. In *post-hoc* analyses we re-estimated the proposed model using a multivariate normal distribution for the random item effects. Our results indicated virtually no correlation among the random effects. However, for many assessments, its reasonable to suspect that a rater who is above average at discriminating on one item may also be above average at discriminating on other items.

Third, having established noninvariance, an important follow-up question examines the extent to which rater characteristics systematically predict noninvariance. For example, do raters with more years of experience demonstrate a greater capacity to discriminate among quality levels? To address this line of inquiries, the proposed model can be further extended to include explanatory components such that random item effects are modeled as a function of fixed rater characteristics through a latent regression framework (De Boeck and Wilson, 2004).

Fourth, the results of our case study suggested that adjustment for persistent differences in severity among raters actually decreased the correspondence between observation and value-added scores. More specifically, when we compared the results of the multilevel model that did not adjust for rater effects at all with that of the cross-classified model with rater severity adjustments (but no random item effects), the correlation between teaching and value-added scores decreased (see Multilevel vs. CC in **Table 6**). These differences could be spurious but they raise questions concerning the value of uniform adjustments for rater severity. In another *post-hoc* analysis, we re-estimated the random item effects cross-classified model (Equation 4) but omitted the overall adjustment for rater severity (γ_r). Our results indicated that absolute fit remained the same but that the correlation between observation and value-added scores increased to 0.20. Again, although the authority of these differences is unknown, these results question the conventional wisdom of including broad sweeping and uniform adjustments for rater severity. Future investigations should examine the fidelity of such adjustments and further consider the efficacy of interactions among the facets. For instance, literature has found that raters function differently across subgroups so that they are more severe within certain subgroups than others.

In conclusion, meaningful comparisons among participants on latent traits in rater-mediated assessments require measurement to be invariant across raters. In many instances, this assumption will be unrealistic. The proposed method offers a flexible alternative that can accommodate measurement noninvariance within multilevel and cross-classified frameworks even when there are no invariant items. Our results suggest the approach is promising and flexible but that it needs more investigation.

ACKNOWLEDGMENTS

The research reported here was supported by grants from the Institute of Education Sciences, U.S. Department of Education (R305C090023), the William T. Foundation and the Spencer Foundation. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

Albert, J. H., and Chib, S. (1993). Bayesian analysis of binary and polytomous response data. *J. Am. Stat. Assoc.* 88, 669–679. doi: 10.1080/01621459.1993.10476321

Asparouhov, T., and Muthén, B. (2010a). *Bayesian Analysis using Mplus: Technical Implementation.* Available online at: http://www.statmodel.com/download/Bayes3.pdf

Asparouhov, T., and Muthén, B. (2010b). *Bayesian Analysis of Latent Variable Models Using Mplus.* Available online at: http://www.statmodel2.com/download/BayesAdvantages18.pdf

Asparouhov, T., and Muthén, B. (2012). *General Random Effect Latent Variable Modeling: Random Subjects, Items, Contexts and Parameters.* Available online at: http://www.statmodel.com/download/NCME12.pdf

Baayen, R. H., Davidson, R. H., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005

Baumgartner, H., and Steenkamp, J. B. E. (2001). Response styles in marketing research: a cross-national investigation. *J. Market. Res.* 38, 143–156. doi: 10.1509/jmkr.38.2.143.18840

Bejar, I. I., Williamson, D. M., and Mislevy, R. J. (2006). "Human scoring," in *Automated Scoring of Complex Tasks in Computer-Based Testing,* eds D. M. Williamson, R. J. Mislevy, and I. I. Bejar (Mahwah, NJ: Lawrence Erlbaum Associates), 49–82.

Bell, C., Gitomer, D., McCaffrey, D., Hamre, B., Pianta, R., and Qi, Y. (2012). An argument approach to observation protocol validity. *Educ. Assess.* 17, 62–87. doi: 10.1080/10627197.2012.715014

Brophy, J. (1986). Teacher influences on student achievement. *Am. Psychol.* 41, 1069–1077. doi: 10.1037/0003-066X.41.10.1069

Carlisle, J., Kelcey, B., and Berebitsky, D. (2013). Teachers' support of students' vocabulary learning during literacy instruction in high poverty elementary schools. *Am. Educ. Res. J.* 50, 1360–1391. doi: 10.3102/0002831213492844

Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. *Am. J. Psychiatry* 126, 884–888.

De Boeck, P. (2008). Random item IRT models. *Psychometrika* 73, 533–559. doi: 10.1007/s11336-008-9092-x

De Boeck, P., and Wilson, M. (eds.). (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* New York, NY: Springer. doi: 10.1007/978-1-4757-3990-9

De Jong, M. G., Steenkamp, J. B. E., and Fox, J. P. (2007). Relaxing measurement invariance in cross−national consumer research using a hierarchical IRT model. *J. Consumer Res.* 34, 260–278. doi: 10.1086/518532

De Jong, M. G., Steenkamp, J. B. E., Fox, J. P., and Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: a global investigation. *J. Market. Res.* 45, 104–115. doi: 10.1509/jmkr.45.1.104

Eckes, T. (2009a). "Many-facet rasch measurement," in *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (Section H),* ed S. Takala (Strasbourg: Council of Europe/Language Policy Division), 4–9.

Eckes, T. (2009b). "On common ground? How raters perceive scoring criteria in oral proficiency testing," in *Tasks and Criteria in Performance Assessment: Proceedings of the 28th Language Testing Research Colloquium,* eds A. Brown and K. Hill (Frankfurt: Lang.), 43–73.

Engelhard, G. (2002). "Monitoring raters in performance assessments," in *Large-Scale Assessment Programs for all Students: Validity, Technical Adequacy, and Implementation,* eds G. Tindal and T. M. Haladyna (Mahwah, NJ: Lawrence Erlbaum Associates, 261–287.

Fox, J. (2010). *Bayesian item Response Modeling: Theory and Applications*. New York, NY: Springer. doi: 10.1007/978-1-4419-0742-4

Fox, J. P. (2007). Multilevel IRT modeling in practice with the package mlirt. *J. Stat. Softw.* 20, 1–16.

Fox, J. P., and Verhagen, A. J. (2010). "Random item effects modeling for cross-national survey data," in *Cross-Cultural Analysis: Methods and Applications,* eds E. Davidov, P. Schmidt and J. Billiet (London: Routledge Academic), 467–488.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis, 2nd Edn.* Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472. doi: 10.1214/ss/1177011136

Gordon, M. (1979). The assessment of impulsivity and mediating behaviors in hyperactive and nonhyperactive boys. *J. Abnorm. Child Psychol.* 7, 317–326. doi: 10.1007/BF00916541

Hansen, M., Lemke, M., and Sorensen, N. (2013). *Combining Multiple Performance Measures: Do Common Approaches Undermine Districts' Personnel Evaluation Systems?* Available online at: http://www.air.org/files/Combining_Multiple_Performance_Measures.pdf

Hickman, J. J., Fu, J., and Hill, H. C. (2012). *Technical Report: Creation and Dissemination of Upper-Elementary Mathematics Assessment Modules*. Priceton, NJ: Eudctional Testing Service.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., et al. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: an exploratory study. *Cogn. Instr.* 26, 430–511. doi: 10.1080/07370000802177235

Hill, H. C., Charalambous, C. Y., and Kraft, M. A. (2012). When rater reliability is not enough: teacher observation systems and a case for the generalizability study. *Educ. Res.* 41, 56–64. doi: 10.3102/0013189X12437203

Holland, P. W., and Wainer, H. (eds.). (1993). *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Jeffreys, H. (1961). *Theory of Probability*. Oxford: Oxford University Press.

Kane, T. J., and Staiger, D. O. (2012). *Gathering Feedback for Teachers: Combining High-Quality Observations with Student Surveys and Achievement Gains.* Available online at: http://www.metproject.org/downloads/MET_Gathering_Feedback_Practioner_Brief.pdf

Lahuis, D. M., and Avis, J. M. (2007). Using multilevel random coefficient modeling to investigate rater effects in performance ratings. *Org. Res. Methods* 10, 97–107. doi: 10.1177/1094428106289394

Linacre, J. M. (1989). *Many-Facet Rasch Measurement*. Chicago, IL: MESA Press.

Linacre, J. M., and Wright, B. D. (2002). Construction of measures from many facet data. *J. Appl. Meas.* 3, 484–509.

Lindley, D. V., and Smith, A. F. (1972). Bayes estimates for the linear model. *J. R. Stat. Soc.* 34, 1–41.

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Messick, S. (1989). "Validity," in *Educational Measurement, 3rd Edn.,* ed R. L. Linn (New York, NY: American Council on Education and Macmillan), 13–104.

Millsap, R. E., and Everson, H. T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Appl. Psychol. Meas.* 17, 297–334. doi: 10.1177/014662169301700401

Muthén, B., and Asparouhov, T. (2013). *BSEM Measurement Invariance Analysis. Mplus Web Notes: No. 17.* Available online at: https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, L. K., and Muthén, B. O. (1998–2012). *Mplus User's Guide, 7th Edn.* Los Angeles, CA: Muthén and Muthén.

Patz, R. J., Junker, B. W., Johnson, M. S., and Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *J. Educ. Behav. Stat.* 27, 341–384. doi: 10.3102/10769986027004341

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edn.* Thousand Oaks, CA: Sage.

Rijmen, F., and Jeon, M. (2013). Fitting an item response theory model with random item effects across groups by a variational approximation method. *Ann. Oper. Res*. 206, 647–662. doi: 10.1007/s10479-012-1181-7

Schochet, P. Z., and Chiang, H. S. (2010). *Error Rates in Measuring Teacher and School Performance Based on Student Test Score Gains.* Available online at: http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Steenkamp, J. B. E., and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *J. Cons. Res.* 25, 78–107. doi: 10.1086/209528

Steinmetz, H. (2013). Analyzing observed composite differences across groups. Is partial measurement invariance enough? *Methodology*, 9, 1–12. doi: 10.1027/1614-2241/a000049

Sudweeks, R. R., Reeve, S., and Bradshaw, W. S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing* 9, 239–261. doi: 10.1016/j.asw.2004.11.001

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Org. Res. Methods* 5, 139–158. doi: 10.1177/1094428102005002001

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

Verhagen, A., and Fox, J. (2013). Bayesian tests of measurement invariance. *Br. J. Math. Stat. Psychol*. 66, 383–401.

Werry, J. S., Sprague, R. L., and Cohen, M. N. (1975). Conners' teacher rating scale for use in drug studies with children-An empirical study. *J. Abnorm. Child Psychol.* 3, 217–229. doi: 10.1007/BF00916752

Wolfe, E. W. (2004). Identifying rater effects using latent trait models. *Psychol. Sci*. 46, 35–51.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 02 June 2014; accepted: 30 November 2014; published online: 23 December 2014.

*Citation: Kelcey B, McGinn D and Hill H (2014) Approximate measurement invariance in cross-classified rater-mediated assessments. Front. Psychol. **5**:1469. doi: 10.3389/fpsyg.2014.01469*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

APPENDIX

EXAMPLE Mplus CODE

```
TITLE: Random item effects;
DATA: FILE IS data.dat;
VARIABLE: NAMES ARE tid rid rich wws cwcm spmmr;
CATEGORICAL =r ich wws cwcm spmmr;
CLUSTER = rid tid;
ANALYSIS: ESTIMATOR = BAYES;
TYPE = CROSSCLASSIFIED RANDOM;
Process=2;

MODEL:
%WITHIN%
s1-s4 |fw by orich* owws cwcm ospmmr;
fw@1;

%BETWEEN TID%
ft BY orich* owws cwcm ospmmr (p1-p4);
s1-s4@0;

%BETWEEN RID%
[s1-s4] (p1-p4);
fw;
```

frontiers in
PSYCHOLOGY

METHODS ARTICLE
published: 08 January 2015
doi: 10.3389/fpsyg.2014.01495

Assessing factorial invariance of two-way rating designs using three-way methods

Pieter M. Kroonenberg *

Department of Child and Family Studies, Institute of Education and Child Studies, Leiden University, Leiden, Netherlands

Edited by:
Alain De Beuckelaer, Radboud University Nijmegen, Netherlands

Reviewed by:
Heungsun Hwang, McGill University, Canada
Pietro Cipresso, IRCCS Istituto Auxologico Italiano, Italy
Wim Krijnen, University of Groningen, Netherlands

****Correspondence:***
Pieter M. Kroonenberg, Department of Child and Family Studies, Institute of Education and Child Studies, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, Netherlands
e-mail: kroonenb@fsw.leidenuniv.nl

Assessing the factorial invariance of two-way rating designs such as ratings of concepts on several scales by different groups can be carried out with three-way models such as the Parafac and Tucker models. By their definitions these models are double-metric factorially invariant. The differences between these models lie in their handling of the links between the concept and scale spaces. These links may consist of unrestricted linking (Tucker2 model), invariant component covariances but variable variances per group and per component (Parafac model), zero covariances and variances different per group but not per component (Replicated Tucker3 model) and strict invariance (Component analysis on the average matrix). This hierarchy of invariant models, and the procedures by which to evaluate the models against each other, is illustrated in some detail with an international data set from attachment theory.

Keywords: stimulus-response data, Tucker3 model, Tucker2 model, Parafac model, three-mode analysis, rating scales, semantic differentials, Ainsworth strange situation

1. INTRODUCTION

Two-way rating designs may consist of, for instance, ratings of concepts on several rating scales. In this paper we tackle the problem of the invariance of the factorial structure of data arising from such designs when the data have been collected from several groups. In particular we will show that three-mode component models are ideally suited to assess factorial invariance for such designs. We will specify a hierarchy of models with increasing restrictions on the parameters resulting in more and more invariant factorial structures across groups.

Because in this paper we are dealing with component models we will use the term "components" rather than "factors," unlessfactors are explicitly indicated. However, to stay within the standard terminology we will use the term *factorial invariance*, rather than *subspace invariance* or *component invariance*. A detailed treatment of the differences between factor analysis and component analysis for two-way data can for instance be found in Widaman (2007).

1.1. FACTORIAL INVARIANCE IN TESTS

Most of the research on factorial invariance assumes that an investigator wants to evaluate whether a test with a particular dimensional structure operates in the same way for different groups, so that the test, or the factors underlying it, can be used for all kinds of groups; a detailed technical exposition of measurement invariance, factorial invariance and their relationship can be found in Meredith (1993). Factorial invariance is typically of interest, for instance, when intelligence tests have been translated into other languages and researchers want to establish whether the translated tests function in the same manner as the original. Alternatively, a researcher may want to know whether a test has the same structure for different groups, say both for regular and for clinical samples.

In a literature survey Vandenberg and Lance (2000, pp. 12–13) synthesized common practices in a list of sequential tests to assess the extent of factorial invariance. The steps in their hierarchy of hypotheses are listed below, but we have listed their first step as the final one, because it is the most restrictive of all invariance schemes, i.e., there is no intergroup variability. Here we present a compact version of their descriptions. Finally, we have added a new first step: Lack of factorial invariance. We need this step later on as a reference point or baseline for our analyses. Note that each next step introduces *additional restrictions* on the parameters of the models.

1. *Lack of invariance*: All groups have different factor patterns.
2. *Configural invariance*: Invariant patterns of factor loadings across groups.
3. *Metric invariance* : Invariant values of factor loadings for like items across groups.

∗a *Scalar invariance*: Invariant intercepts of like items regressions on the factor.

∗b *Unique variances invariance*: Invariant unique variances of like items across groups.

c *Invariant factor variances*: Invariant factor variances across groups.

4. *Invariant factor covariance matrices*: Invariant factor covariance matrices across groups.

∗d *Invariant factor means*: Invariant factor means across groups.

5. *Strict invariance*: Invariant factor means and covariance matrices across groups.

The hierarchy is primarily based on investigations using factor analysis within the context of structural equation modeling with and without estimation of the factor means. This means that it contains concepts and parameters characteristic of such models, such as unique variances, factor means and intercepts of regressions of items on factors. In this paper such concepts do not play a role, because our proposals are based on component analysis. In the sequel, the starred steps are therefore excluded for the following reasons: (*a, *d) all scales will be centered across concepts for each group (see below), so that means and factor means do not enter into the models; (*b) the concept of unique variances does not play a role in component analysis. Note that when referring to Step 5, "Strict invariance," we will assume only that the covariance matrices are equal across groups, again because the means have already been removed by centering.

The major analytical techniques for establishing the increasingly stricter types of invariance have primarily been structural equation modeling and item response theory as is evident in this special issue. In the hierarchy of hypotheses about factorial invariance it is implied that the models are nested, so that they can be evaluated, or in the context of structural equation models, tested against each other. This means that an a priori choice has to be made about the factor model itself: How many factors and which items are to be regressed on which factors. Therefore, a two-factor model may be invariant in a different way than a three-factor model for the same data. In this paper we will concentrate on series of both two-factor and three-factor models, but we will not attempt to make detailed comparisons between the two series.

Regarding the component models in this paper, comparisons between models are primarily based on the error sums of squares in relation to their degrees of freedom. These degrees of freedom are calculated as the number of data points minus the number of parameters to be estimated (N_{parm}) where the means subtracted during the centering of the data are also counted as parameters. Details and formulas for calculating the degrees of freedom for three-way models can for instance be found in Kroonenberg (2008, Section 8.4, p. 177ff).

1.2. TWO-WAY RATING DESIGNS

In psychology a specific kind of measurement design is commonly used, i.e., a *two-way rating design* in which concepts are judged on scales by a number of judges such as in Osgood's classical semantic differential design (Osgood et al., 1957). Alternative two-way rating designs generate stimulus-response data or situation-scale data. Characteristic for the designs is that a subject has to judge to what extent a particular scale or variable pertains to a particular concept or situation. For instance, in a study by Murakami and Kroonenberg (2003), a student had to judge the characteristics of the 24 preludes of Chopin on a number of scales. As example, the student had to indicate whether a prelude of Chopin (concept) is tempestuous or tranquil (scale). Another example, which will be our guiding explanatory case, is the two-way design in which a person with a multiple personality in each personality was asked to judge on a number of scales to what extent a number of concepts pertained to her personal situation. For instance, to what extent she considered her doctor to be good or bad (Osgood and Luria, 1954). The aim in their study was to see whether each personality (Eve White, Eve Black and Jane; each measured twice) used the scales in the same way to rate the concepts.

Yet another kind of two-way rating data results from a design in which for several situations the mean characteristics of groups rather than of individual subjects are described by means of a number of variables. For our detailed example we analyzed a collection of two-way data sets consisting of episodes by variables obtained from several different countries. The data were collected using the Strange Situation, a procedure within the attachment theory paradigm (Ainsworth et al., 1978) (see Section 3).

A two-way rating design seems comparable to multitrait-multimethod (MTMM) designs where the traits and the methods mostly form a fully-crossed design for the response variables. An important difference with the MTMM design is that the two-way rating design is more like a two-way (concept×scale) analysis-of-variance design with the intensity or strength of the judgment by a personality as the response variable.

1.3. TWO-WAY RATING DESIGNS AND THREE-WAY DATA

Two-way rating designs produce three-way data because they consist of three ways, i.e., concepts, scales and groups or individuals. For a more detailed discussion of such *three-way rating data* arising from two-way rating designs see Kroonenberg (2008, Chapter 14). As far as we have been able to trace, there is no or hardly no explicit literature on the topic of factorial invariance for two-way rating designs, and with this paper we aim to fill this gap. In particular, our aim is to look for both a consensus structure about the relations between the concepts and scales (i.e., invariance over groups) and for group differences, i.e., deviations from invariance. Even though we will primarily focus on the situation with a limited number of groups or individuals, also larger numbers can be analyzed. The emphasis in the present paper is an exploratory one, even though the comparative evaluation of different aspects of factorial invariance using fit measures is a central concern. However, the sizes and relevance of these differences have to be evaluated subjectively both by comparing fit/degrees of freedom ratios and by looking at substantive relevance and interpretability. Formal statistical testing is not part of the procedure.

1.4. INVARIANCE IN TWO-WAY RATING DESIGNS

A problem for the invariance analysis of two-way rating designs is that there are often only a limited amount of judges or groups rather than large samples from a population so that there is no clear stochastic element in the data. The judges or groups need to be treated as another fixed factor in the analysis-of-variance sense, so that we really have a three-way design of concepts × scales × groups or concepts × scales × individuals. Even apart from the extremely small samples, this lack of stochastics in two-way rating designs makes using confirmatory factor analysis for testing invariance within the standard structural equation modeling context virtually impossible. Therefore, we propose to seek recourse to variants of component analysis, but it should be noted that the procedures discussed in this paper can handle large random samples as well.

Factorial invariance for two-way rating designs is cast here in a non-stochastic component framework in which we have

separate component spaces for the scales and the concepts. This has a disadvantage because components are generally not in themselves meaningful quantities but only maximum variance directions in the component space. What are invariant are the subspaces spanned by the components, rather than the components themselves. Therefore, we cannot automatically assume that the components themselves have intrinsic meaning like factors in confirmatory common factor analysis.

Only in some very specific models, such as the Parafac models which have unique solutions (see below), the components can validly be said to have intrinsic meaning. This will limit the kinds of invariances we can consider. Thus, generally we will have to discuss the invariance of subspaces across groups rather than the invariance of the components themselves. As already indicated in the introduction rather than refer to *subspace invariance* or *componential invariance*, we will use the standard term *factorial invariance*.

The two central questions in two-way rating designs are (1) how to define factorial invariance and (2) how to evaluate it. In contrast with the standard situation of assessing whether factorial invariance exists for a particular test across groups, in a two-way design one has to deal with the fact that *groups* or *individuals* use the rating scales to judge concepts. A definition of factorial invariance in this case must include three aspects of the data: (1) the component space or structure of the scales; (2) the component space or structure of the concepts and (3) the way the concepts (or the concept components) and the scales (or scale components) are linked for each group. The consideration of three different aspects of factorial invariance makes the situation for two-way rating designs fundamentally different from the standard situation. Both because of the design and the fact that we are dealing with component spaces rather than factors, makes that the Vandenberg and Lance steps have to be reformulated.

1.4.1. Preprocessing

Variances of components in standard component analysis are represented by the eigenvalues. Whether they are actual variances or merely corrected or uncorrected sums of squares depends on the preprocessing, i.e., centering and normalization of the data. Standardization is more or less automatically carried out in regular component analysis but in two-way rating designs there are several options for preprocessing. Each option has different consequences for the data to be assessed for invariance, because it influences which part of the data is analyzed (see e.g., Kroonenberg, 2008, Chapter 6). To avoid such complications we will ignore the influence of preprocessing in this paper, and we will use the terms sums-of-squares and variances indiscriminately.

1.5. INVARIANCE HIERARCHY

When adapting the steps in the invariance hierarchy for two-way rating designs, we will assume from the start that we are attempting to approximate the centered data with lower-rank component spaces for the concepts and for the scales. This is in contrast with confirmatory factor analysis where covariance matrices are approximated.

Given the definition of a component, i.e., a linear combination of the original variables, any component is always present in a data set with the same variables given its coefficients; a property called *perfect congruence*; for a detailed discussion of this property see Ten Berge (1986a,b). What is generally different in different data sets with the same variables is the amount of variance explained by the components in each group. When it is not the full component space that is under consideration but only a limited number of (maximum variance) components, these group component spaces can be spanned by different linear combinations of the variables, so that component spaces of different groups may even be orthogonal to each other. The maximum variance components of one group, may account for very little variability in another group.

1.5.1. Step 1. Lack of invariance

The most extreme form of lack of invariance is that each group has its own low-dimensional subspace. For two-way designs we take as our starting point the separate analyses of the group data without imposing any restrictions on the component subspaces other than considering a limited number of components, the same number for each group. The fitted sum of squares of the groups together, the *combined fit*, is calculated by summing their individual fitted sums of squares.

1.5.2. Step 2. Configural invariance

Because every component returns in each data set with the same variables, i.e., components are always perfectly congruent across groups, configural invariance is not a limiting restriction in component analysis and is automatically true. Thus, it cannot be used as a limiting concept in a hierarchy of models, even though in different groups the same components may account for different amounts of variance and have different correlations.

1.5.3. Step 3. Metric invariance

Of the models used to inspect factorial invariance, metric invariance is part of their definition. Thus, the component spaces (for the concepts and scales) specified in the models are such that the component coefficients are identical across groups. Three models can be used to investigate metric invariance. They have either (3a) an invariant concept component space, (3b) an invariant scale space or (3c) both. Metric invariance can be compared with a total lack of invariance by comparing the metric-invariant model fit with the *combined fit*. In addition, the metric invariant space can be compared with the separate spaces of the groups, for instance via Procrustes techniques (see, for instance, Gower and Dijksterhuis, 2004); see also Section 4.

For the component models under consideration we will use the terms *links* and *interactions* to indicate the parameters which link the concepts and scales components. The links are contained in a so-called *core array* $\mathbf{H}$ (see **Figure 1**). For each group this array contains a slice, $\mathbf{H}_k$, with the group's links between the components of the scales and the concepts. If both the concept and the scale space are orthogonal, the sizes of these links are the square roots of variation accounted for by the components. The invariance of the factor covariance matrices across groups translates into the equality of the core slices $\mathbf{H}_k$ for $k = 1, \cdots, K$.

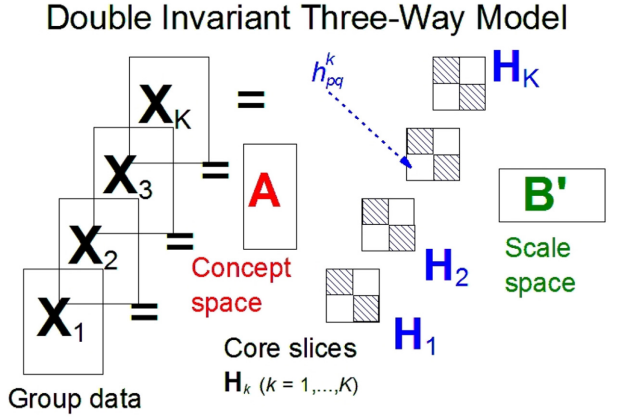


FIGURE 1 | A general three-mode model for two-way rating designs. A = metric invariant concept space; B = metric invariant scale space; $H_k = (h^k_{pq})$ = core slice for the k^{th} group; h^k_{pq} is the link between the pth component of A and the qth component of B.

1.5.4. Step 4. Invariant component covariance matrices or core slices

As no common three-way models have restrictions on the variances without restrictions on the covariances, such models will not be discussed here; see Harshman and Lundy (1984) for detailed considerations about this issue. We will, however, consider (4a) models with invariant covariances (off-diagonal elements of the core slices) for all groups but with different variances (diagonal elements of the core slices). Even more restricted are models in which (4b) the invariant scale and/or concept components are uncorrelated in all groups.

1.5.5. Step 5. (Weighted) strict invariance

The equality of the covariance matrices in Vandenberg and Lance's Step 5 translates into the equality of the centered data matrices of the groups. Such an equality implies equality of random errors which is of course nonsensical. However, a further tightening of the invariance in Step 4 is achieved in Step (5a) by restricting the slices of the core array to be identical, apart from a size coefficient (in the following referred to as a *weight*). Finally, the strictest factorial invariance situation is created in Step (5b) by specifying that also the weights are invariant across groups. In that case the structure of the scales and the concepts, as well as their linkages, are identical in all groups.

1.6. RELATED RESEARCH

Thus, for the two-way rating design the investigation of invariance is concentrated on the linkages between the invariant components for all groups. The discussion of the hierarchy of increasingly invariant three-mode models in this paper is strongly related to the hierarchy of three-mode models for fully-crossed raw data (Kiers, 1991). In addition, a similar hierarchy can be found in connection with simultaneous component analysis of covariance and correlation matrices (Timmerman and Kiers, 2003). However, in those papers the concept of factorial invariance is not the focus of the investigation nor is the emphasis on two-way rating data.

2. MODELING FACTORIAL INVARIANCE

This section deals with three-way models for analysing data two-way rating designs. These models have as a common characteristic that the scale space and the concept space are invariant for all groups. However, they differ in the nature of the linkages between concept and space components. The models in Step 3a and 3b have *metric invariance* in one mode and all other models are characterized by *double-metric invariance*.

2.1. MODELS FOR TWO-WAY RATING DESIGNS

Table 1 provides an overview of appropriate models, together with listing the nature of their invariances. To discuss these models in some detail we need some notation. **A** and **B** indicate the $I \times P$ invariant concept space and the $J \times Q$ invariant scale space, with P and Q the number of components, respectively. A subscript k indicates that a particular matrix belongs to the kth of K groups or levels of the third way; for instance, $\mathbf{X}_k$ is the concept $\times$ scale data matrix of the kth group. $\mathbf{H}_k = (h^k_{ss})$ is the linkage matrix for the concept and the scale components for the kth group, $\mathbf{D}_k$ is a diagonal matrix of links used in the SVD as well as in the Parafac model. In the next section we will discuss these models in detail and indicate how they embody factorial invariance. As indicated in **Table 1** the Tucker2 model in principle allows for different numbers of components for the scales and the concepts, but as it is the only three-way model in **Table 1** for which this is the case, we will assume in the following that $S = P = Q$, i.e., that the numbers of components for the two spaces are the same throughout, so that **A** has size $I \times S$ and **B** has size $J \times S$.

2.2. STEP 1: SINGULAR VALUE DECOMPOSITION PER GROUP

The singular value decomposition (SVD) is the motor of many multivariate techniques. For any $\mathbf{X}_k$ it may be written as:

$$\mathbf{X}_k = \mathbf{A}_k\mathbf{D}_k\mathbf{B}'_k + \mathbf{E}_k = \hat{\mathbf{X}}_k + \mathbf{E}_k \quad k = 1, \cdots, K \qquad (1)$$

where for the SVD to have the form in Equation (1), the concept spaces $\mathbf{A}_k$ and scale spaces $\mathbf{B}_k$ have to have orthogonal components and the linkage matrices $\mathbf{D}_k$ have to be diagonal. The $\mathbf{E}_k$ contain the errors of approximation. $\hat{\mathbf{X}}_k = \mathbf{A}_k\mathbf{D}_k\mathbf{B}'_k$, and $\mathbf{E}_k = 0$ if all components are used. We will refer to the collection of independent analyses for each group as the *separate-analyses model* with abbreviations SVD_2 and SVD_3 for the two- and three-component models, respectively.

Thus, each data matrix $\mathbf{X}_k$ has its own decomposition as in Equation (1), and this decomposition is unrelated to that of any of the other data matrices. The total variance of a group k is equal to the sum of the squares of the singular values d^k_{ss} that make up the diagonal of $\mathbf{D}_k$ in the full decomposition, i.e., $\text{SS(Total)}_k = \sum_k d^k_{ss}$. Adding the SS(Total)_k of the groups gives the total amount of variance of the groups indicated by SS(Total). In general, we will use only a limited number of components, here either 2 or 3. The components (columns) of $\mathbf{A}_k$ and $\mathbf{B}_k$ successively account for the largest amount of variance so that, given the dimensionality, the components for the concepts and those

Table 1 | Models for two-way rating designs and their invariance.

Model	Concepts	Scale	*P* = *Q*?	Interaction	Abbreviation
STEP 1: LACK OF INVARIANCE					
SVD per group	-	-	yes	no explicit invariance restrictions	SVD_*s*
STEP 3: METRIC INVARIANCE					
Tucker1 - concepts invariant	x	-	no	concept space invariant; *single metric invariance*	T1A_*s*
Tucker1 - scales invariant	-	x	no	scale space invariant *single metric invariance*	T1B_*s*
Tucker2	x	x	no	concept and scale spaces invariant; *double-metric invariance*	T2_*ss*
STEP 4: INVARIANT COMPONENT COVARIANCES					
Parafac	x	x	yes	+ component covariances invariant; variances free	PF*s*
Parafac - Orthogonal	x	x	yes	+ component covariances invariant; variances free; components orthogonal for one or both ways	PF*s*_Orth
STEP 5: (WEIGHTED) STRICT INVARIANCE					
Tucker3 - Free	x	x	yes	metric invariance of orthogonal components variances invariant; group weights unrestricted	T3_*ss*1
Tucker3 - Fixed	x	x	yes	+ group weights fixed and constant	T3_*ss*1Fixed

x = invariant; S-S = not invariant; SVD, Singular Value Decomposition; P = number of concept components; Q = number of scale components; s = 2 or 3 number of components; ss1 = the first two ways have s components, the third way 1 component.

for the scales span the subspaces with the highest variance. Thus, we can use this variance accounted for, $\text{SS(Fit)}_{\text{separate}}$, as an upper bound for the variance accounted for from any other model given the number of components. If the SS(Fit) is the fit for a common model for all K groups, then if $\text{SS(Fit)}_{\text{model}} \simeq \text{SS(Fit)}_{\text{separate}}$ the component space(s) are invariant. However, if there is a sizeable difference, the invariance restrictions on the common model are in doubt. We may also investigate *group invariance* by comparing the fitted variance of a particular group SS(Fit)_k with the similar quantity calculated via the parameter estimates from one of the fitted models. Given the number of components, this will provide information on which groups fit well and which groups do not and are thus not invariant with respect to the other groups.

2.3. STEP 3A AND STEP 3B: SINGLE METRIC INVARIANCE - TUCKER1 MODELS

The first step into imposing restrictions on the solutions to investigate possible invariance is to demand that either the concept spaces can be properly represented by a single space (i.e., for all k the concept spaces are equal: $\mathbf{A}_k = \mathbf{A}$), or that for all k the scale spaces are equal: $\mathbf{B}_k = \mathbf{B}$ if there are s components. This can be investigated with the Tucker1 model, here referred to as Tucker1A (or T1A_*s*) for concept space equality and Tucker1B (T1B_*s*) for scale space equality. Metric invariance exists for the concepts if

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}\mathbf{B}_k' + \mathbf{E}_k \quad k = 1, \cdots, K. \tag{2}$$

Thus, there is a single orthogonal concept space for all k and separate scale spaces for each group. Metric invariance exists for the scales if

$$\mathbf{X}_k = \mathbf{A}_k\breve{\mathbf{D}}\mathbf{B}' + \mathbf{E}_k \quad k = 1, \cdots, K. \tag{3}$$

Thus, there is a single orthogonal scale space for all k and separate concept spaces for each group.

To compute the parameters, the three-way array is first converted to a two-way matrix of (Groups × Scales) by Concepts or (Groups × Concepts) by Scales, and these matrices are then subjected to a SVD. Note that the resulting $\mathbf{A}_k$ and $\mathbf{B}_k$ are no longer orthogonal because they are parts of a single orthogonal matrix of left and right singular vectors, respectively. We may compare the fitted variance of these models $\text{SS(Fit)}_{\text{model}}$ with the combined results of the separate SVDs, $\text{SS(Fit)}_{\text{separate}}$, to investigate the metric invariance of either the concept or the scale spaces. However, it seems a bit odd to have an invariant concept space without having an invariant scale space, so we will not include the Tucker1A model further in our deliberations.

2.4. STEP 3C: DOUBLE-METRIC INVARIANCE - TUCKER2 MODEL

The next step in imposing invariance is to require *double-metric invariance*, i.e., for all k and given a number of components s both $\mathbf{A}_k = \mathbf{A}$ and $\mathbf{B}_k = \mathbf{B}$, where both matrices orthogonal. Furthermore, the group linkage matrices $\mathbf{H}_k$ are unrestricted and thus in general not diagonal. The model equation for the Tucker2 model (Tucker, 1972), as the model is commonly known (see Kroonenberg, 2008, Section 4.5.2) becomes

$$\mathbf{X}_k = \mathbf{A}\mathbf{H}_k\mathbf{B}' + \mathbf{E}_k \qquad k = 1, \cdots, K. \tag{4}$$

In other words, the metric invariance is present on both the concept space and the scale space, and the only differences between the groups can occur in the K interaction or linkage matrices, $\mathbf{H}_k$. The linkages matrices $\mathbf{H}_k$ have sizes $S \times S$, where S is the number of components for both the scale and the concept spaces. An element h_{pq}^k of $\mathbf{H}_k$ represents the link between the pth component of the concepts and the $q$the component of the scales for the kth group. So apart from their error terms, the variability between the groups lies in the strengths of their links between the concept and scale components or the sizes of the h_{pq}^k.

We can again compare the fitted variance of these models $SS(Fit)_{model}$ with the combined results of the separate SVDs, $SS(Fit)_{separate}$, to investigate the double-metric invariance. Similarly we can make comparisons at group level.

2.5. STEP 4: DOUBLE-METRIC INVARIANCE WITH INVARIANT CORRELATIONS - PARAFAC MODEL

By requiring $\mathbf{H}_k = \mathbf{C}_k$, where the latter are diagonal matrices, and dropping the orthogonality restriction on the component spaces, we get the standard Parafac model with s components (PFs) which is a double-metric invariant model with as its model equation

$$\mathbf{X}_k = \mathbf{A}\mathbf{C}_k\mathbf{B}' + \mathbf{E}_k \qquad k = 1, \cdots, K. \tag{5}$$

The model can also be written by filling the rows of a $K \times S$ matrix $\tilde{\mathbf{C}}$ with the diagonals of the $\mathbf{C}_k$, i.e., $\tilde{c}_{ks} = c^k_{ss}$ $k = 1, \cdots, K$. In that case $\tilde{\mathbf{C}}$ is considered a component matrix and is normalized like $\mathbf{A}$ and $\mathbf{B}$, i.e., the lengths of the components in all three matrices are equal to one. The sizes of the S components are then contained in a diagonal matrix $\mathbf{D} = (d_{ss})$. However, for this paper we will stick with the $\mathbf{C}_k$.

Harshman (1970) that has shown this model implies that the groups have the same correlations between the components, which is a further imposition of factorial invariance. When at least one of the component matrices is orthogonal the d^2_{ss} are the variances of the S components.

One can even impose further restrictions on the components and so make the invariance even stricter by reintroducing orthonormality, non-negativity, or unimodality on one or both component matrices (see, e.g., Bro and Sidiropoulos, 1998).

Compared to other three-way models, Parafac models have a special characteristic in that their parameters are uniquely determined under rather mild conditions. This implies that the parameters in Equation (5) cannot by altered, for instance by rotation, without lowering the fit. The consequence is that the model has the *parallel proportional profile* property; (see Cattell and Cattell, 1955; Harshman, 1970; Harshman and Lundy, 1984). The only lack of invariance in these models consists of different strengths of the links between the concepts and scales, i.e., the c^k_{ss} vary between the groups. From the parallel proportional profile property and the uniqueness of the models it is the components themselves, not only the subspaces they span which are invariant; see Harshman (1970) or Harshman and Lundy (1984).

2.6. STEP 5: STRICTLY INVARIANT MODELS - TUCKER3 MODELS

To study factorial invariance with even more restrictions, we can demand that for each k $c^k_{ss} = c^k d_{ss}$. In other words the weights for the components are invariant across groups apart from a group weight c^k.

$$\mathbf{X}_k = \mathbf{A}(c^k\mathbf{D})\mathbf{B}' + \mathbf{E}_k = c^k(\mathbf{A}\mathbf{D}\mathbf{B}') + \mathbf{E}_k \qquad k = 1, \cdots, K. \tag{6}$$

This model equals a simplified version of the full Tucker3 model (Tucker, 1966), and has been referred to as the *Replicated PCA* model by Van IJzendoorn and Kroonenberg (1990) and *Weighted PCA* by Krijnen and Kiers (1995). The only variable parts are the weights c^k for the group applicable to both components, and the error terms $\mathbf{E}_k$. In other words, all groups have the same concept and scale spaces and the orthogonal components of each way are linked such that each concept component is linked exclusively to a particular scale component. The part between brackets has the form of a SVD valid for all groups. The only differences between the groups are their weights, c^k. This is in contrast with the Parafac model where each group has different link weights for the concept and scales component combinations, i.e., the c^k_{ss} are different for each group k and each pair of components s.

The ultimate invariant model is that in which we assume that all c_k are all equal with weight $\bar{c} = \sqrt{1/K}$, which is computationally equivalent to first averaging over groups and then carrying out a SVD on the average data matrix $\bar{\mathbf{X}}$, i.e.,

$$\mathbf{X}_k = \bar{c}(\mathbf{A}\mathbf{D}\mathbf{B}') + \mathbf{E}_\mathbf{k} \qquad k = 1, \cdots, K. \tag{7}$$

Thus, in this case the only variable parts are the error terms and we may speak of *strict invariance*. We could reduce even further the number of parameters by specifying further restrictions on the concept and scale component spaces (see Takane et al., 1995), but this will not be considered here.

2.7. SUMMARY EVALUATING INVARIANCE

The conclusion from the above subsections is that one can define a hierarchy of models with an ever increasing number of parameters which are invariant over groups. By comparing the models with each other and with the combined separate analyses, it becomes possible to evaluate which models still provide an adequate fit to the data compared to separate analyses, and hence which type of invariance can be safely adopted. The two leading types of information for this purpose are the overall fitted variance and the fitted variance of each group.

In order to carry out model comparisons the number of parameters estimated for each of the models is determined. The models are compared by constructing a variant of the three-mode scree plot, in which the fitted sum of squares are plotted against the number of parameters estimated (see Section 3.3). Details on how to calculate the number of parameters can be found in Kroonenberg (2008, Section 8.4).

3. EXAMPLE: THE STRANGE SITUATION ACROSS THE WORLD

3.1. RESEARCH DESIGN

Attachment between adults, especially mothers, and infants is a lively research area—(see Cassidy and Shaver, 1999, 2008. Three types of bonds between adults and infants are generally considered: Avoidant attached, Securely-attached, and Resistant/Ambivalent attached, indicated by the letters A, B, and C, respectively. Here we will only look at attachment bonds with mothers, but those with other adults, especially other caregivers, have also been investigated (see, e.g., Sagi et al., 1985). The measurement procedure consists of a series of episodes of approximately 3 min, during each of which the infant is in a standardized room together with the mother (M), the stranger (S), both (MS), or alone (A); the episodes are the following: M1, MS2, S3, M4, A5, S6, M7. The idea is to increase the stress on the infant, especially by introducing the stranger and leaving the child alone, so that

the attachment relationship between mother and infant is put to the test. During the episodes, except when the infant is alone (A5), five core variables of an infant's reaction to an adult are measured: Proximity seeking, Contact maintaining, Avoidance, Resistance, and Distance interaction.

3.1.1. Strange situation data set

The data set under consideration consists of 11 samples: US-Belsky (USBel), US-Thompson (USTho), Germany-Berlin (GerBe), Germany-Bielefeld (GerBi), Israel-Kibbutz (IsrKi), Israel-City (IsrCi), Japan-Miyake (JapMi), Japan-Takahashi (JapTa), Netherlands-Younger infants (NLYng), Netherlands-Older infants (NLOld) and Sweden (Swed). The data set was put together by Sagi and Lewkowicz, and in their publication (Sagi and Lewkowicz, 1987) they supply full details of the origins of the different samples. For each of the samples the original investigators independently determined the infants' type of attachment. Earlier analyses can be found in Sagi and Lewkowicz (1987) and Kroonenberg and Van IJzendoorn (1987).

3.1.2. Invariance

The research question for this paper is whether the structure of the scales and that of the episodes, as well as the way these components are linked, are invariant across samples. The more parameters in the models are invariant, the more evidence this presents that the Strange Situation is a valid procedure across countries and researchers. For this example we only examine the average scores of the samples securely attached infants (B). These samples were chosen because each contained a sufficient number of B infants to make the average scores reliable. Thus, the two-way rating design consists of 7 episodes by 5 scales for 11 samples. This three-way data set was subjected to the models described above and their fit measures were compared.

3.2. RESULTS: THREE-WAY ANALYSIS OF VARIANCE

To acquire an initial perspective on the differences between samples, we carried out a three-way analysis of variance of the Strange Situation data. For this analysis the response variable was considered to be intensity of a reaction, and the Three-Ways were conceived as fixed factors in the ANOVA sense. This view is feasible because the samples are not exchangeable or drawn from a population. Moreover, it is the individual differences between the samples which are the focus of the analysis. Furthermore, the scales all had the same range from 1 to 7, so that averaging across scales is feasible and interpretable.

Table 2 shows that the largest variability is between scales, indicating that the scale scores of the infant-mother dyads are effective in differentiating between behaviors across samples and episodes. On the other hand, the sample variability is comparatively very small (2.2% of the total), indicating that the investigating factorial invariance is a worthwhile exercise. This is confirmed by the size of the episode × scale interaction compared to the interactions involving samples. Finally, the residuals (or the three-way interaction) only take up 7.5% of the total variability.

Parallel with standard component analysis, before the three-way analyses the data were centered but not normalized.

Table 2 | Three-Way analysis of variance (with a single observation per cell).

Source	SS	% SS(Total)	*df*	MS	*F*
MAIN EFFECTS					
Episodes	79.3	16.4%	6	13.2	88.1
Scales	211.4	43.6%	4	52.8	324.0
Samples	10.7	2.2%	10	1.1	6.5
TWO-WAY INTERACTIONS					
Episodes × Scales	105.2	21.7%	24	4.4	26.9
Episodes × Samples	9.9	2.0%	60	0.2	1.0
Scales × Samples	29.5	6.1%	40	0.7	4.5
THREE-WAY INTERACTION					
Residuals	39.1	8.1%	240	0.2	
TOTAL	485.1				

df = degrees of freedom; MS = Mean Sum of squares.

Normalization was not deemed necessary because all the scales had the same range. Moreover, scales with more variability should be allowed to have more influence on the analysis than scales with little variability.

With respect to centering, the common type of centering for three-way rating scale data (averaging across the concepts) was used, i.e., $\tilde{x}_{ijk} = (x_{ijk} - \bar{x}_{.jk})$. In other words, the scale means for each sample k were removed. In general, centering across samples is undesirable because it will eliminate the consensus configuration of the scales and concepts from the three-way analysis. Thus, due to this type of centering the means of the scales for each of the samples were not included in the invariance analysis, but depending on the purpose of a study, these means can be analyzed for invariance separately.

3.3. RESULTS: INVESTIGATING TYPE OF INVARIANCE VIA MODEL FIT

Because the procedure outlined for assessing factorial invariance for two-way rating designs is an exploratory one, deciding on the degree of invariance is a substantive and subjective matter, of course based on numerical information. **Table 3** provides the information on the series of more and more restricted, and hence more invariant, models. Any additional restriction on the parameters is going to incur a certain amount of additional loss compared to the separate analyses. However, the question is whether the decrease in fit can be acceptable, given that by restricting the number of parameters interpretability is enhanced. It is less useful to compare the two-component models with the three-component models, because they have different starting points, i.e., different separate solutions. Therefore, it seems best to first decide on the number of components one wants to use to model the data, and only after that to investigate the invariance. This is incidentally also the standard practice in structural equation modeling. Of course, one may come to the conclusion that a two-component model is more, or less, invariant than a three-component model and vice versa.

In **Table 3** we see that the most restrictive models are the Tucker3 models with a constant component for the samples (T3-221Fixed and T3-331Fixed), i.e., the strictly invariant models. At the other extreme the individual three-component SVDs are not much use in terms of data reduction, because the model for

each sample has only three degrees of freedom, and the rank of the centered data matrices is at most four. From a data-analytic point of view, it is doubtful whether a model with unrestricted three-component solutions for the separate samples is really useful because the three components fit about 97% of the total variability.

To decide upon the most appropriate model for these data, and thus on the extent of the invariance, it is useful to construct a variant of the three-mode deviance plot of the fitted sums of squares vs. the number of parameters (**Figure 2**); see Kroonenberg, 2008, Section 8.5. The models with two components and those with three have been connected by part of a convex hull. Models on a convex hull are generally preferred to the models inside such a hull because of their more favorable SS(Fit)/N_{Parms} ratios. It is preferable to consider only models on or very close to the convex hull; the PF3-Orth model is less attractive because there are models with more favorable ratios (PF3 and T3-331) in the neighborhood. The more horizontal a hull, the more a model on the right is a good alternative for the models to the left on the hull, because the decrease in the number of parameters (i.e., increase

Table 3 | Overall sums-of-squares for the Strange Situation data.

	Model	Abbreviation	SS(Fit)	SSS(Fit)	*df*	N_{Parms}
TWO-COMPONENT SOLUTIONS						
Step 1	SVD per group	(SVD_2)	205.78	0.89	110	275
Step 3	Tucker1 - Scales invariant	(T1B_2)	187.94	0.81	227	158
	Tucker2	(T2_22)	177.78	0.76	270	115
Step 4	Parafac	(PF2)	173.54	0.74	288	97
	Parafac - Orthogonal scale components	(PF2_Orth)	173.36	0.74	290	95
Step 5	Tucker3 + Variable weights	(T3_221)	169.18	0.72	302	83
	Tucker3 + Fixed weights	(T3_221Fixed)	164.77	0.71	312	73
THREE-COMPONENT SOLUTIONS						
Step 1	SVD per group	(SVD_3)	227.05	0.97	33	352
Step 3	Tucker1 - Scales invariant	(T1B_3)	219.53	0.94	154	231
	Tucker2	(T2_33)	206.57	0.88	213	172
Step 4	Parafac	(PF3)	200.40	0.86	267	118
	Parafac - Orthogonal scale components	(PF3_Orth)	195.26	0.84	273	112
Step 5	Tucker3 + Variable weights	(T3_331)	185.43	0.79	299	86
	Tucker3 + Fixed weights	(T3_331Fixed)	181.23	0.78	309	76

N_{Parms} = Number of parameters (includes 55 removed means due to centering); The Total Sum of Squares of the centered data: SS(Tot) = 233; SSS(Fit) = SS(Fit)/SS(Tot); df = degrees of freedom = number of data points (I × J × K = 385) − N_{Parms}.

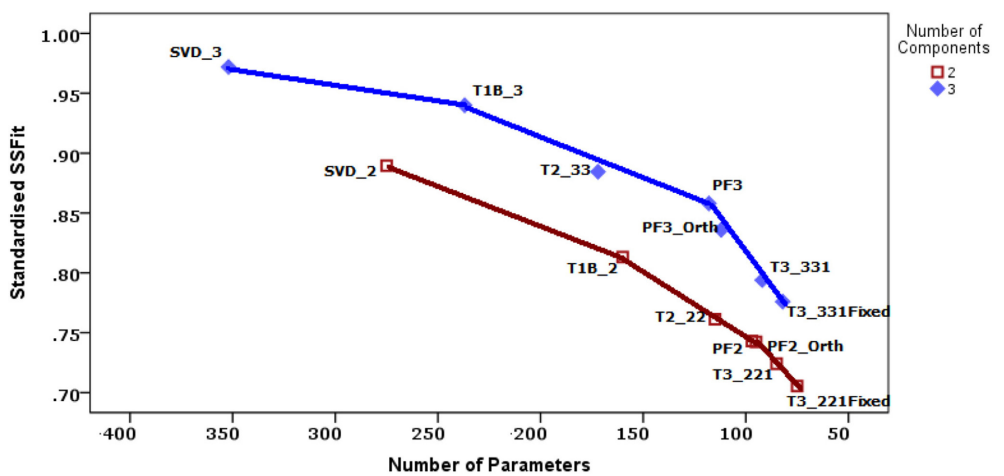


FIGURE 2 | Model comparisons. The two-component and three-component models are connected by separate convex hulls. The horizontal axis is reversed because the investigation starts with the individual models. *Legend*: SVD_*s* = separate SVDs with *s* components; T1B_*s* = Tucker1 model with *s* components; T2_*ss* = Tucker2 model with *s* components for way 1 and 2; PF*s* = Parafac model with *s* components; PF*s*_Orth = Parafac model with *s* orthogonal scale components; T3_*ss*1 = Tucker3 model with *s* components for way 1 and 2 and 1 component for way 3; T3_*ss*1Fixed = T3_*ss*1 with a fixed value for way 3.

in the *df*) does not seriously decrease the fitted sum of squares. In contrast, the steeper the hull turns downward for the next model to the right, the less attractive the model, because there is a large loss in fitted sum-of squares for only a limited decrease in parameters. Note that a smaller number of parameters increases power and potentially simplifies interpretability.

For the Strange Situation data we see in **Figure 2** that for the three-component models the convex hull declines slowly at first, and a steeper downturn is observed only for the Tucker3 models, so that the Parafac model with three components seems a good choice. The choice for a two-component model is less clear. The relationship between the SS(Fit) and the number of parameters is nearly linear. Again the Parafac models (PF2 and PF2-Orth) seem to be the best choice, and even though the orthogonal variant is marginally better, we decided to opt for the regular Parafac model. With respect to factorial invariance, the Parafac models incorporate invariant concept and scale spaces, and the correlations between scale components are constant over samples. The appropriateness of the Parafac models suggests that there is a considerable double-metric factorial invariance across the samples, only the size of the variances is different.

3.4. RESULTS: NON-INVARIANT SAMPLES

For three-way models with double-metric invariance which are not necessarily invariant with respect to their links, we can compute the model fit for each sample. These fit measures can then be compared with the separate-analyses model to determine whether overall lack of interaction invariance is due to specific samples or whether differences are present between all samples.

3.4.1. Differences in proportional fit of samples.

For selected two-component models we calculated the proportional residual sums-of-squares PrSS(Res$_k$) for each sample and connected these values per model in **Figure 3**. In the figure we have arranged the samples such that the lack of fit is increasing for the two-component Parafac model.

The solid line for the PrSS(Res) represent SVDs of the separate samples. We see that their PrSS(Res) fluctuate around the average value drawn as a horizontal line. In other words, a two-component SVD have about the same fitted sums of squares in all samples, but their concept spaces and their scale spaces are not necessarily equal.

In the case of strict model invariance all lines would be more or less horizontal because the lack of fit would be equal for all samples. This is not the case here. The relative difference in fit varies between the solutions for the separate samples and those of the models displayed in the figure. Thus, for the US samples on the left-hand side of the figure the metric invariant subspaces for the concepts and the scales are more alike to their own separate spaces than to the subspaces for the younger Dutch sample and Israel-City sample on the right-hand side. In particular, the PrSS(Res) for the two US samples is around 0.10 while it is around 0.30 for the younger Dutch and the Israel City sample.

All three metric invariant models displayed in **Figure 3** show more or less the same pattern with an increasing loss of fit from left to right. Given that the models are more or less equivalent, we may choose to interpret the most restricted and thus most invariant model, i.e., the T3-221 or PF2 models. **Figure 3** shows that the most right-hand samples fit marginally better, which is consistent with our earlier choice for this model. The Parafac model allows the components per sample to have a common oblique orientation with separate weights (c_{ss}^k) for the links between these common components. In this data set the younger Dutch sample and the Israel City sample need further investigation, because it is their configurations that are deviating most from the common pattern.

3.4.2. Differences in strengths of links between concept and scale spaces across samples

In **Figure 4** we have plotted the link strengths c_{ss}^k between the concept and scale components from the Parafac model with two components. The solid line represents the strengths of the

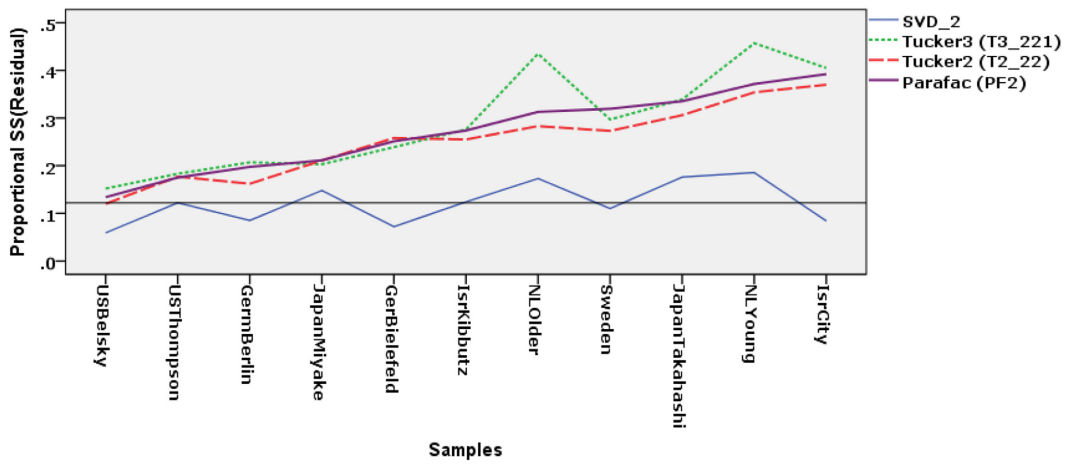


FIGURE 3 | Proportional Residual sums of squares per sample for four two-component models: Separate-analysis model (SVD-2), Parafac model (PF2), Tucker2 model (T2-22), and Tucker3 model (T3-221). The samples are ordered on their fit based on the Parafac model with two components.

Legend: US-Belsky (USBel), US-Thompson (USTho), Germany-Berlin (GerBe), Germany-Bielefeld (GerBi), Israel-Kibbutz (IsrKi), Israel-City (IsrCi), Japan-Miyake (JapMi), Japan-Takahashi (JapTa), Netherlands-Younger infants (NLYng), Netherlands-Older infants (NLOld) and Sweden (Swed).

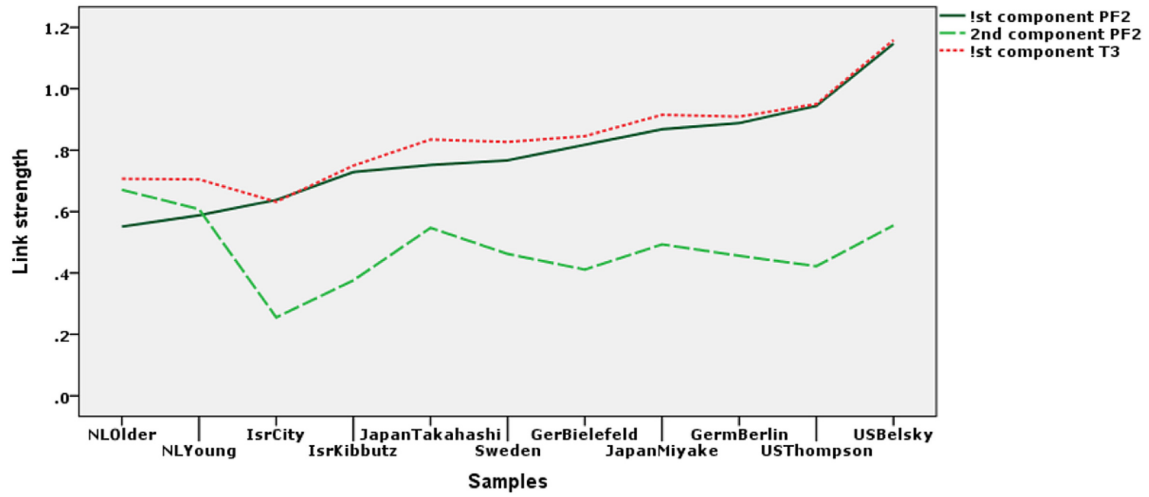


FIGURE 4 | The strength of the Parafac component links (c^k_{ss}) for each of the samples in principal coordinates. The dotted line represents the weight or strength of the links from the T3_221 model (c^k) in principal coordinates. For the abbreviations of the samples names see **Figure 3**.

links between the first components, c^k_{11} and the dotted line the strength of the links for the second components, c^k_{22}. To provide a proper comparison these parameters have been depicted in principal coordinates. The third, dashed, line represents the weight parameter for each group according to the T3-221 model, c_k; also in principal coordinates. The samples have been ordered so that the values for the first components, c^k_{11}, are increasing monotonically. The figure shows that c^k_{11} and the c^k are almost equal, but that there is a small compensation of the c^k for the absence of the links for the second components c^k_{22}. Thus, the choice between the models should take into account whether the fluctuations of the c^k_{22} are interpretable. At the same time the differences in the c^k_{22} point to where we should look for lack of invariance.

If we want to find out what exactly are the differences between the samples, we have to explicitly compare the invariant concept and scale spaces with the separate sample spaces. Thus, this analysis could be extended to find the causes of the differences by examining the Tucker1 model for scales (T1B), and possibly the Tucker1 model for concepts (T1A), to assess whether it is the scale space or the concept space which is not invariant. We will not pursue this here. The procedure described above should primarily be seen as a proof of concept, rather than a detailed analysis of a particular case (see, however, the Appendix for a more substantive interpretation).

4. RESULTS: AN ADDITIONAL APPROACH TOWARD ASSESSING INVARIANCE

In a paper comparing Japanese and Australian children in the way they show respect to adults, Kroonenberg and Kashima (1997) tackled assessing invariance in a different way, even if they did not explicitly refer to factorial invariance. The children were given a questionnaire in which they had to indicate both to what extent they did *show* a number of respectful behaviors (greet, help, stick up for, etc.) toward a number of adults (father, mother, teacher, etc.), and to what extent they felt they *should* do so. This resulted in a 5 (adults) × 7 (behaviors) × 4 (groups; Australian do, Australian should, Japanese do, Japanese should) three-way data set. Apart from a complete three-way analysis, the invariance was also assessed by first carrying out separate analyses for each of the four groups, and then using the adult space and/or the behavior space of one group as a restriction for the solution of another group. Essentially, of course, this is a cross-validating procedure, checking to what extent the parameter estimates in one group can also explain the variability in another group, or to what extent the two groups had invariant subspaces. However, one may equally see this as a procedure for establishing invariance. This procedure was referred to as *external analysis* by Van der Kloot and Kroonenberg (1985), because externally determined values for the parameters were used in fitting a particular data set.

For the Strange Situation data, this procedure could be used to investigate to what extent the separate solution of a sample is similar to that of another sample. In particular, the nature of the difference of the Dutch sample with respect to the other samples could be a focus of further analysis.

5. CONCLUSION

In this paper we have presented an approach toward assessing factorial invariance in two-way rating designs such as stimulus-response and semantic differential designs. Such designs generate fully-crossed three-way data which can be analyzed by three-way component models. True three-way models like the Parafac and Tucker models and their variants already incorporate various aspects of factorial invariance, in particular the double-metric invariance of the concept and scale spaces. The models vary in how they treat the relationships or links between the components. A hierarchy of models with increasing factorial invariance is outlined, running from no invariance for separate SVDs for each group, via single metric invariance for Tucker1 models, double-metric invariance for Tucker2 models, double-metric invariance and correlational invariance of Parafac models, to strict invariance for a very restricted Tucker3 model.

These models, and hence the nature of the invariance, can be assessed and compared via deviance plots showing the sum of squares of fit against the degrees of freedom. By connecting the relevant models by convex hulls in the plot, a comparative evaluation can be made and an appropriate model can be selected. Moreover, information supplied by the three-way analysis can be used to assess which group is more deviant from the invariant solution, and what the nature of such differences are.

The descriptive approach toward model selection, rather than using a formal testing paradigm, has been shown to work well for the example presented here. Data from a multinational collection of Strange Situation sessions (Sagi and Lewkowicz, 1987) were analyzed to demonstrate the effectiveness and usefulness of the model hierarchy for two-way rating data.

By investigating data from two-way rating designs we have extended the concept of factorial invariances beyond its standard definition. The future will have to show to what extent this extension is going to make an impact on the research on factorial invariance. For the present it seems that using the conceptualization presented here and the proposed hierarchy of three-way models, can shed light on differences and similarities between the invariance in two-way rating designs.

ACKNOWLEDGMENTS

Our thanks go to Avi Sagi, Kathleen Lewkowicz, and Marinus van IJzendoorn for making the international Strange Situation data available for reanalysis.

REFERENCES

Ainsworth, M. D. S., Blehar, M. C., Waters, E., and Wall, S. (1978). *Patterns of Attachment. A Psychological Study of the Strange Situation*. Hillsdale, NJ: Erlbaum.

Bro, R., and Sidiropoulos, N. D. (1998). Least squares algorithms under unimodality and non-negativity constraints. *J. Chemometr.* 12, 223–247. doi: 10.1002/(SICI)1099-128X(199807/08)12:4<223::AID-CEM511>3.0.CO;2-2

Cassidy, J., and Shaver, P. R., (eds.). (1999). *Handbook of Attachment. Theory, Research, and Clinical Applications*. New York, NY: The Guilford Press.

Cassidy, J., and Shaver, P. R., (eds.). (2008). *Handbook of Attachment, 2nd Edn. Theory, Research, and Clinical Applications.* New York, NY: The Guilford Press.

Cattell, R. B., and Cattell, A. K. S. (1955). Factor rotation for proportional profiles: analytical solution and an example. *Br. J. Stat. Psychol.* 8, 83–92. doi: 10.1111/j.2044-8317.1955.tb00323.x

Gower, J. C., and Dijksterhuis, G. B. (2004). *Procrustes Problems*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780198510581.001.0001

Harshman, R. A. (1970). Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Work. Pap. Phon.* 16, 1–84.

Harshman, R. A., and Lundy, M. E. (1984). "The PARAFAC model for three-way factor analysis and multidimensional scaling," in *Research Methods for Multimode Data Analysis*, eds H. G. Law, C. W. Snyder Jr., J. A. Hattie, and R. P. McDonald (New York, NY: Praeger), 122–215.

Kiers, H. A. L. (1991). Hierarchical relations among three-way methods, *Psychometrika* 56, 449–470. doi: 10.1007/BF02294485

Krijnen, W. P., and Kiers, H. A. L. (1995). An efficient algorithm for weighted PCA. *Comput. Stat.* 10, 299–306.

Kroonenberg, P. M. (2008). *Applied Multiway Data Analysis*. Hoboken NJ: Wiley. doi: 10.1002/9780470238004

Kroonenberg, P. M., and Kashima, Y. (1997). Rules in context. A three-mode principal component analysis of Mann et al.'s data on cross-cultural differences in respect for others. *J. Cross Cult. Psychol.* 28, 463–480. doi: 10.1177/0022022197284005

Kroonenberg, P. M., and Van IJzendoorn, M. H. (1987). "Exploring children's behavior in the Strange Situation," in *Attachment in Social Networks. Contributions to the Bowlby-Ainsworth Attachment Theory*, eds L. W. C. Tavecchio and M. H. Van IJzendoorn (Amsterdam: Elsevier Science), 379–425. doi: 10.1016/S0166-4115(08)61080-8

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Murakami, T., and Kroonenberg, P. M. (2003). Individual differences in semantic differential studies and their analysis by three-mode models. *Multivariate Behav. Res.* 38, 87–96. doi: 10.1207/S15327906MBR3802_5

Osgood, C. E., and Luria, Z. (1954). A blind analysis of a case of multiple personality. *J. Abnorm. Soc. Psychol.* 49, 579–591. doi: 10.1037/h0054362

Osgood, C. E., Suci, G., and Tannenbaum, P. (1957). *The Measurement Of Meaning* Urbana, IL: University of Illinois Press.

Sagi, A., Lamb, M., Lewkowicz, K., Shoham, R., Dvir, R., and Estes, D. (1985). Security of infant-mother, -father and -metapelet attachments among kibbutz-reared Israeli children. *Monogr. Soc. Res. Child Dev.* 50, 257–275. doi: 10.2307/3333837

Sagi, A., and Lewkowicz, K. S. (1987). "A cross-cultural evaluation of attachment research," in *Attachment in Social Networks. Contributions to the Bowlby-Ainsworth Attachment Theory*, eds L. W. C. Tavecchio and M. H. van IJzendoorn (Amsterdam: Elsevier Science), 427–459. doi: 10.1016/S0166-4115(08)61081-X

Takane, Y., Kiers, H. A. L., and De Leeuw, J. (1995). Component analysis with different sets of contraints on different dimensions. *Psychometrika* 60, 259–280. doi: 10.1007/BF02301416

Ten Berge, J. M. F. (1986a). Some relationships between descriptive comparisons of components from different studies. *Multivariate Behav. Res.* 21, 29–40. doi: 10.1207/s15327906mbr2101_2

Ten Berge, J. M. F. (1986b). Rotation to perfect congruence and the cross-validation of component weights across populations. *Multivariate Behav. Res.* 21, 41–64. doi: 10.1207/s15327906mbr2101_3

Timmerman, M., and Kiers, H. A. L. (2003). Four simultaneous component models for the analysis of multivariate time series from more than one subject to model intraindividual and interindividual differences. *Psychometrika* 68, 105–121. doi: 10.1007/BF02296656

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika* 31, 279–311. doi: 10.1007/BF02289464

Tucker, L. R. (1972). Relations between multidimensional scaling and three-mode factor analysis. *Psychometrika* 37, 3–27. doi: 10.1007/BF02291410

Van der Kloot, W. A., and Kroonenberg, P. M. (1985). External analysis with three-mode principal component models. *Psychometrika* 50, 479–494. doi: 10.1007/BF02296265

Van IJzendoorn, M. H., and Kroonenberg, P. M. (1990). Cross-cultural consistency of coding the Strange Situation. *Infant Behav. Dev.* 13, 469–485. doi: 10.1016/0163-6383(90)90017-3

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Widaman, K. F. (2007). Common factors versus components, principals and principles, errors and misconceptions, in *Factor Analysis at 100. Historical Developments and Future Directions* eds R. Cudeck and R. C. MacCallum (Mahwah, NJ: Lawrence Erlbaum), 177–203.

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 28 February 2014; accepted: 04 December 2014; published online: 08 January 2015.

*Citation: Kroonenberg PM (2015) Assessing factorial invariance of two-way rating designs using three-way methods. Front. Psychol. **5**:1495. doi: 10.3389/fpsyg.2014.01495*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

APPENDIX

This appendix is presented here to offer some idea of the substantive outcomes of the invariance analysis of the Strange Situation.

For the Parafac model with two components, **Figure A1** shows the normalized components of the three modes in two panels, both for the first and the second component as well as the strength of the links between them.

If there was complete strict invariance, the samples would have been superimposed in both panels at the value($\bar{c}d_{ss}$ (the T3-221Fixed model). If there would have been weighted strict invariance, the rank order and spacing of the samples for each of the components would have been equal i.e., at the values $c^k d_{ss}$ (the T3-221 model). As the figure shows, neither of these options was realized in the present data set, so that we must conclude that a double-metric invariant model (PF2) is the most restricted or invariant model that can be obtained.

The variances (or link strengths) of the components are $d_{11} = 11.5$ and $d_{22} = 6.4$, respectively (see Equation 5), so that the ratio of their importance in reconstructing the model is 1.8. Thus, the differences between the samples with respect to link strengths of the first components are about twice as large as those for the second components.

A.1. FIRST COMPONENT

The left-hand panel of **Figure A1** shows that securely attached (B) children show increasing Proximity seeking and Contact maintaining during the Mother episodes of the Strange Situation, as is evident from the increasingly higher coefficients on the first component. Seeking closeness to the mother is indicative of increasing stress during the procedure, which the B children try to alleviate by showing more and more proximity to the mother, i.e., showing a stronger secure attachment behavior. Treating the stranger with suspicion by staying at a distance is evident in the Stranger episodes; the coefficients remain negative but less so in S6 than in S3. Children's suspicion is decreasing slightly during the procedure but it is never absent. The other three scales all hover around zero, indicating that these behaviors of the children are not related to the two behaviors mentioned first. The US securely-attached children show the described patterns to the largest extent and the Dutch children the least. It is interesting to see that samples from the same country are generally close together with the largest difference between the two Israeli samples.

A.2. SECOND COMPONENT

The second components in the right-hand panel of **Figure A1** describe mainly the avoidance, resistance and distance interaction behaviors toward the stranger, or stranger wariness. Such behavior is not typically present in the first two episodes but is present to a limited extent in the other episodes except for Episode 6, when it is the Stranger who returns rather than the Mother after the child has been alone in the fifth episode. With respect to the mother the situation is more complicated. There is a clear contrast between the earlier and later episodes, in that negative behavior toward the mother is not present in the beginning, but the children show a certain reserve when mother and child are reunited after the child has been alone with the stranger (Episodes M4 and M7).

These patterns are strongest in the Dutch and Japanese samples, as well as Belsky's US sample. Again, samples from the same country are generally close together except for the US samples.

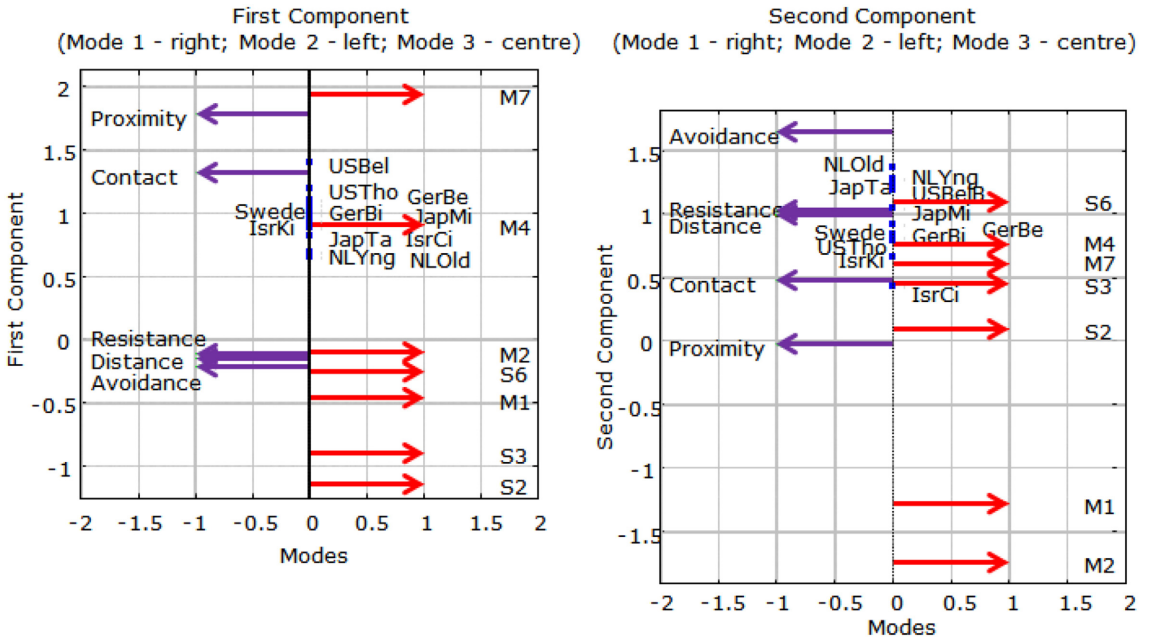


FIGURE A1 | Two-component Parafac analysis: Invariant components plus link strength of samples. M*i* = Mother episode *i*; S*i* = Stranger episode *i*. For sample abbreviations, see **Figure 3**.

A.3. INVARIANCE CONCLUSION WITH RESPECT TO THE B-CHILDREN

The original research question was whether the structure behind the data from the two-way rating design for the secure B-children was the same across samples. The double-metric invariance embodied in the well-fitting two-component Parafac model indicates that this is indeed the case. However, the samples show a lack of invariance with respect to importance of the linkage between the components of the episodes and the scales. This difference is primarily a matter of relative importance of the two components. The first component embodies the increasingly secure attachment behaviors (Proximity seeking and Contact maintaining) over episodes, which is stronger in the US samples, especially in contrast with the Dutch samples. The second component represents 'stranger wariness' (Avoidance, Resistance and Distance interaction) which is especially strong in Episode 6, when the stranger rather than the mother returns, after the child as been alone in the fifth episode. To a lesser degree it also represents reservation toward the mother in the reunion episodes. These patterns are especially strong in the Dutch and Japanese samples and Belsky's US one.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 18 November 2014
doi: 10.3389/fpsyg.2014.01304

The experience of traumatic events disrupts the measurement invariance of a posttraumatic stress scale

Miriam J. J. Lommen[1], Rens van de Schoot[2,3]* and Iris M. Engelhard[4]

[1] *Experimental Psychology, University of Oxford, Oxford, UK*
[2] *Method and Statistics, Utrecht University, Utrecht, Netherlands*
[3] *Optentia Research Program, Faculty of Humanities, North-West University, Vanderbijlpark, South Africa*
[4] *Clinical and Health Psychology, Utrecht University, Utrecht, Netherlands*

Edited by:
Peter Schmidt, University of Giessen, Germany

Reviewed by:
Kenn Konstabel, National Institute for Health Development, Estonia
Tenko Raykov, Michigan State University, USA

****Correspondence:***
Rens van de Schoot, Department of Method and Statistics, Utrecht University, PO Box 80.140, 3508 TC Utrecht, Netherlands
e-mail: a.g.j.vandeschoot@uu.nl

Studies that include multiple assessments of a particular instrument within the same population are based on the presumption that this instrument measures the same construct over time. But what if the meaning of the construct changes over time due to one's experiences? For example, the experience of a traumatic event can influence one's view of the world, others, and self, and may disrupt the stability of a questionnaire measuring posttraumatic stress symptoms (i.e., it may affect the interpretation of items). Nevertheless, assessments before *and* after such a traumatic event are crucial to study longitudinal development of posttraumatic stress symptoms. In this study, we examined measurement invariance of posttraumatic stress symptoms in a sample of Dutch soldiers before and after they went on deployment to Afghanistan ($N = 249$). Results showed that the underlying measurement model before deployment was different from the measurement model after deployment due to invariant item thresholds. These results were replicated in a sample of soldiers deployed to Iraq ($N = 305$). Since the lack of measurement invariance was due to instability of the majority of the items, it seems reasonable to conclude that the underlying construct of PSS is unstable over time if war-zone related traumatic events occur in between measurements. From a statistical point of view, the scores over time cannot be compared when there is a lack of measurement invariance. The main message of this paper is that researchers working with posttraumatic stress questionnaires in longitudinal studies should not take measurement invariance for granted, but should use pre- and post-symptom scores as different constructs for each time point in the analysis.

Keywords: measurement invariance, posttraumatic stress disorder, trauma, threshold instability, multiple assessments

INTRODUCTION

Questionnaires are often used at different time points to assess mean or individual change over time. For example, a questionnaire to assess posttraumatic stress symptoms can be rated at different time points after a traumatic event to study the course of problematic responses. Although statisticians have stressed the importance of testing measurement invariance when comparing latent mean scores over time (e.g., Byrne et al., 1989; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000), the assumption that factor loadings and intercepts (or thresholds when dealing with dichotomous or categorical scores instead of continuous scores) of the underlying items are equal over time often seems to be taken for granted. By comparing latent mean scores over time, we aim to capture *true* latent score changes (i.e., alpha change; Brown, 2006). However, in case of measurement non-invariance, increases or decreases in latent mean scores may also reflect changes in the construct itself (gamma change) or changes in the measurement proportions of the indicators (beta change). Therefore, it is important that factor loadings and intercepts are "measurement invariant" to claim *true* latent score change over time and to avoid bias in the parameter estimates (Guenole, 2014). But what should one do in case of measurement non-invariance? Is it then still possible to draw meaningful conclusions or should mean scores over time not be compared? In this article we discuss a measure that, from a theoretical perspective, is expected to lack measurement invariance. In such cases the solutions of establishing partial invariance (Byrne et al., 1989) or approximate invariance (van de Schoot et al., 2013; Muthén, 2014) are not a valid solution. We will test for measurement invariance in two samples, and investigate causes of measurement non-invariance and interpretations of the results in this situation.

THE CASE OF THEORETICAL MEASUREMENT NON-INVARIANCE

The experience of a traumatic event can lead to psychological distress, which may manifest as posttraumatic stress disorder (PTSD). PTSD is characterized by re-experiencing symptoms (e.g., intrusions or nightmares related to the event), avoidance of reminders of the event, negative cognitions and mood, and hyperarousal symptoms (e.g., sleep and concentration problems; APA, 2013). One way to check the presence of PTSD symptoms

is by using self-report questionnaires. Although it is often not possible to include a pre-trauma assessment of symptomatology, several prospective longitudinal studies, typically in military or firefighter samples, have done this and showed that PTSD symptoms after a traumatic event may partially be explained by symptoms endorsed at baseline (e.g., Engelhard et al., 2007b; Rona et al., 2009; Vasterling et al., 2010; Rademaker et al., 2011; van Zuiden et al., 2011; Berntsen et al., 2012; Bonanno et al., 2012; Franz et al., 2013; Lommen et al., 2013, 2014). High scores at baseline could represent symptoms that are not exclusively related to PTSD (e.g., sleep or concentration problems, negative mood; Engelhard et al., 2009b), or they may reflect already existing PTSD symptoms resulting from earlier traumatic experiences. So when prospectively studying, for instance, predictors for the development of PTSD symptoms, it seems useful to take symptoms that were already present before trauma into account.

However, it may be hypothesized that the experience of a traumatic event[1] (APA, 2013) can actually change the way items of the questionnaire are interpreted. That is, after experiencing a traumatic event, the probability of answering "yes" to a specific questions may increase or decrease (gamma change), and the relative importance of questions may change (beta change).

Consider, for example, soldiers who complete a questionnaire for PTSD symptoms before and after deployment. Before deployment, soldiers may be instructed to rate the items in reference to a recent event that made them feel especially upset or distressed, in reference to a distressing event that bothered them the most in the last month, or without reference to a specific event. After deployment, the soldier may be instructed to fill out the questionnaire with respect to most distressing event during the recent deployment, or to rate the symptoms without reference to a specific event. Before deployment, the presence of symptoms could relate to a range of events or stressors. After deployment, the symptoms are likely a reaction to the warzone experiences in which life-threatening situations are experienced or witnessed, like being shot at, being exposed to the explosion of an improvised explosive device (IED), or having to help with the removal of human remains. Such experiences can drastically change one's view on the world, like perceiving the world as a dangerous place, and one's evaluative reactions (e.g., Foa and Rothbaum, 1998; Ehlers and Clark, 2000; Engelhard et al., 2009a, 2011). Moreover, common posttraumatic symptoms like having unexpected, distressing thoughts about the event, nightmares, and sleeping problems can be negatively interpreted and may lead to a change in the soldier's view on his/her self, such as "I am a weak person," or "My reactions since the event mean that I am going crazy" (Foa et al., 1999). The question that arises is whether it is realistic to expect measurement invariance for the situation as described here.

In sum, assessing levels of PTSD symptoms at baseline as well as after the traumatic events is essential to model the development of PTSD symptoms, but may be statistically problematic at the same time because of expected measurement non-invariance.

THIS STUDY

In the current study, we tested measurement invariance in two datasets that were part of two larger prospective studies about resilience and vulnerability factors involved in PTSD symptoms (see Lommen et al., 2013 for sample 1, and Engelhard et al., 2007b for sample 2). Using Sample 1, we investigated the source of the measurement non-invariance, including the effect of the presence or absence of prior deployment experiences. Arguably, those with prior deployment experiences are more likely to fill out the questionnaire with regard to deployment related traumatic experiences at both time points. Expecting measurement invariance may therefore be specifically unrealistic for the group without prior deployment experience. Sample 2 was used to test whether the results of sample 1 would be replicated. Finally, solutions for dealing with non-invariant data will be discussed.

MATERIAL AND METHODS

Sample 1 consisted of 249 Dutch soldiers [Task Force Uruzgan (TFU) 11], who completed the Dutch version (Engelhard et al., 2007a) of the Posttraumatic Symptom Scale—Self Report (PSS; Foa et al., 1993) about 2 months before their 4-month deployment to Afghanistan ($N = 249$), and about 2 months after their return home ($n = 241$). The PSS is a self-report questionnaire with 17 items that represent the 17 symptoms of PTSD according to the DSM-IV (American Psychiatric Association, 2000), which includes (a) re-experiencing symptoms, such as intrusions, flashbacks, and nightmares (b) avoidance symptoms (e.g., avoidance of reminders of the traumatic event) and numbing, and (c) hyperarousal symptoms, such as hypervigilance, sleep disturbances, and concentration problems. Before their deployment, participants were asked to rate the questions with respect to their most aversive life-event that troubles them the most in the last month. After deployment, participants were instructed to complete the PSS with respect to their deployment-related event(s) that troubled them the most in the last month. Items were rated on a 0 (*not at all*) to 3 (*almost always*) scale. For convenience, scores were dichotomized into 0 *(symptom absent)* to 1 *(symptom present)* for the analyses.

Sample 2 consisted of 305 Dutch soldiers, derived from a larger study in which 481 soldiers were included [stabilization Force Iraq (SFIR) 3, 4, and 5; Engelhard et al., 2007b]. Since only SFIR 3 and 5 were asked to complete the PSS before their deployment, these two groups were included in this study ($N = 310$). Only soldiers who completed the PSS at least at one of the two time points were included in this study ($n = 305$). Before their deployment to Iraq, 291 soldiers filled out the PSS, and 242 soldiers completed the PSS about 5 months after their return home.

At the post-deployment assessment, both samples completed a Dutch version of the Potentially Traumatizing Events Scale (PTES;

[1]Exposure to actual or threatened death, serious injury or sexual violation. The exposure must result from one or more of the following scenarios, in which the individual:

- directly experiences the traumatic event;
- witnesses the traumatic event in person;
- learns that the traumatic event occurred to a close family member or close friend (with the actual or threatened death being either violent or accidental); or
- experiences first-hand repeated or extreme exposure to aversive details of the traumatic event (not through media, pictures, television or movies unless work-related).

Maguen et al., 2004), which assessed the frequency of exposure to war-zone related stressors. For sample 1, the questionnaire was adjusted to the situation in Afghanistan, resulting in 24 stressors (cf. Lommen et al., 2013). For sample 2, the questionnaire was adjusted to the situation in Iraq, resulting in 22 stressors (cf. Engelhard and van den Hout, 2007). Participants indicated whether they had experienced each stressor, and the negative impact (no, mild, moderate, or severe).

Participation was strictly voluntary without financial compensation. Both prospective projects were approved by the Institutional Review Board of Maastricht University.

DATA ANALYSIS

Analyses were conducted with Mplus 7.11 (Muthén and Muthén, 2010). First, using Sample 1, two confirmatory factor analyses (CFA) for the PSS at the two time points were assessed. Second, measurement invariance was tested, as suggested by Raykov et al. (2012) by comparing the model fit of four competing, but nested, models: the unconstrained CFA model (factor loadings and thresholds of the latent variable were freely estimated), the CFA model with threshold invariance (constrained thresholds), the CFA model with loading invariance (constrained factor loadings), and the CFA model with scalar invariance (constrained factor loadings and thresholds). The tests for determining measurement invariance were repeated for Sample 2 to investigate whether the results for Sample 1 could be replicated. Third, to investigate whether the measurement invariance test would be different for soldiers with and without prior deployment experiences, the previous step was repeated for these two groups separately. Fourth, to gain insight in the source of potential measurement non-invariance we applied two methods: (1) differences in factor loadings and thresholds were tested using a Wald test; and (2) we employed the method of Raykov et al. (2013). For the first method we used the loading invariance model and tested each pair of thresholds using the MODEL TEST option in Mplus. This procedure resulted in 17 Wald tests. For the second method, of Raykov et al., we first tested the chi square difference (using the DIFFTEST option of Mplus) between the scalar model and 17 models (17 items) where one pair of thresholds was left unconstrained at a time (Method 2A). This resulted in 17 chi square difference tests. If all tests in comparison to the scalar model are non-significant, then measurement invariance holds. If some tests are significant whereas others are not, we can conclude that partial invariance holds and we know which items are causing the non-invariance. Since the CFA models indicated that the loading invariance model showed the best fit (with thresholds freely estimated), we also computed the chi difference tests between the loading invariance model and 17 models where one set of thresholds was constrained (Method 2B). This latter procedure is a replication of the first method, with the MODEL TEST option, but this time with chi square values instead of Wald tests. The two methods (i.e., 2A and 2B) can be considered as the forward and backward methods of sequential regression analyses and will probably result in slightly different solutions just like with sequential analyses.

For the Raykov method we applied the Benjamini-Hochberg multiple testing procedure as described in Raykov et al. (2013). That is, we calculated a corrected alpha value, indicated by l in the tables. The p-values of the chi square difference tests should then be smaller than l instead of the default alpha of.05. After computing the chi square differences, the resulting p-values are ordered from small to large and for each row a different l value is computed. For more details, how to compute l and syntax-examples we refer to Raykov et al. (2013). In the appendix of our paper we provide our Mplus syntax for the final model of method 1 (all other syntax files can be found at the website of the second author: www.rensvandeschoot.com) and in the footnote of **Table 3** we provide the code for obtaining l.

The root mean square error of approximation (RMSEA, Steiger, 1990), comparative fix index (CFI; Bentler, 1990), and Tucker-Lewis index (TLI; Tucker and Lewis, 1973) were used to evaluate model fit. RMSEA values of <0.08, CFI, and TLI values of >0.90 were considered to reflect adequate model fit (see Kline, 2010 for an overview of fit statistics). To compare models, we used Chi square difference test, Akaike Information Criterion (AIC; Akaike, 1981) and Bayesian Information Criterion (BIC; Schwarz, 1978) values.

RESULTS

EXPERIENCED EVENTS ON DEPLOYMENT

The most commonly experienced deployment-related events in all samples (TFU 11 of sample 1, SFIR 3 and SFIR 5 of sample 2) were "Going on patrols or performing other dangerous duties" (90–94%), "Fear of being ambushed or attacked" (65–95%), and "Fear of having unit fired on" (61–95%). Amongst those events that participants rated as having a moderate to severe negative impact were "Being informed of a Dutch soldier who got killed" (21–51%), "Witnessing an explosion" (9–25%), "Seeing dead or injured Dutch soldiers" (0–24%), and "Having to aid in the removal of human remains" (0–13%).

SAMPLE 1

CFA models including the latent variable PSS loading on 17 indicators showed acceptable model fit at both time points [before deployment: $\chi^2_{(119)} = 175.027$, $p < 0.001$, RMSEA (90% CI) = 0.044 (0.029–0.058), CFI = 0.961, TLI = 0.955; after deployment: $\chi^2_{(119)} = 175.237$, RMSEA (90% CI) = 0.044 (0.029–0.058), CFI = 0.921, TLI = 0.909]. **Table 1** presents an overview of the fit indices used to evaluate the CFA-models including PSS at both time points. The CFA including PSS at both time points with freely estimated factor loadings and the CFA with loading invariance showed acceptable model fit. The model fit of the unconstrained CFA was better according to the chi square difference test, CFI, TLI, and RMSEA, but the CFA with loading invariance (see Appendix 1 for Mplus syntax of model statement) was better according to the AIC and BIC. The CFA that imposed threshold invariance and the one imposing scalar invariance both showed unacceptable model fit. The results of all fit indices indicate that the measurement non-invariance has mainly to do with the instability of the thresholds over time.

SAMPLE 2

Similar to sample 1, the CFA models including the latent variable PSS in sample 2 showed acceptable model fit at both time points [before deployment: $\chi^2_{(119)} = 160.476$, $p = 0.007$, RMSEA (90% CI) = 0.035 (0.019–0.048), CFI = 0.941, TLI = 0.933; after

Table 1 | Model fit information for CFA including PSS before and after deployment in sample 1 and 2.

	χ^2(df)	CFI	TLI	RMSEA (90% CI)	AIC	BIC
SAMPLE 1						
Unconstrained	640.821 (526)	0.924	0.919	0.030 (0.020–0.037)	5974.361	6217.065
Threshold invariance	751.535 (543)	0.862	0.857	0.039 (0.032–0.046)	6034.422	6217.330
Loading invariance	674.540 (543)	0.913	0.910	0.031 (0.023–0.039)	5965.915	6148.823
Scalar invariance	772.401 (560)	0.859	0.859	0.039 (0.032–0.046)	6218.945	6342.056
SAMPLE 2						
Unconstrained	630.235 (526)	0.961	0.959	0.025 (0.017–0.033)	6639.398	6896.100
Threshold invariance	763.777 (543)	0.918	0.915	0.037 (0.030–0.042)	6715.873	6909.330
Loading invariance	618.640 (543)	0.972	0.971	0.021 (0.011–0.029)	6621.558	6815.014
Scalar invariance	726.491 (560)	0.938	0.938	0.031 (0.024–0.037)	6830.930	6961.140

AIC and BIC through MLR, rest: WLSMV.

deployment: $\chi^2_{(119)}$=219.654, RMSEA (90% CI) = 0.059 (0.047–0.071), CFI = 0.963, TLI = 0.957]. Although in this sample all CFA models with varying constrains showed acceptable model fit, AIC and BIC were lowest for the loading invariance model (see **Table 1**). Again, the measurement non-invariance seems to arise from instability of the thresholds.

PRIOR DEPLOYMENT EXPERIENCE

It could be argued that measurement non-invariance would be driven by those participants who have not been deployed before, because they may refer to different types of stressors before and after this particular deployment when rating the items. For those participants who have been deployed before, the meaning of the construct might have already changed with the experience of the prior deployment. Therefore we tested measurement invariance in the group with (56.63 and 41.64% in Sample 1 and 2, respectively) and without prior deployment experience separately. Nevertheless, based on AIC/BIC comparison, the results showed a similar pattern for both groups, suggesting that threshold instability underlies measurement non-invariance in our samples, regardless of the presence or absence of prior deployment experience. The results can be found in the online available supplementary materials.

THRESHOLD INSTABILITY

To gain insight in the instability of the thresholds for both samples, we explored the difference in thresholds for each item between the two time points. For descriptive purposes, the threshold before deployment was subtracted from the threshold after deployment difference to define threshold difference for each item. The threshold represents the mean score on the latent variable that is related to the "turning point" where an item is rated as present instead of not present. Thus, a positive difference score means that compared to the PSS mean score before deployment, a higher PSS mean score was needed to rate an item as present after deployment. Threshold values and difference scores are presented in **Table 2**.

The first method we used to test for threshold differences is to compute a Wald test whether, for each item, the threshold after deployment significantly increased or decreased compared to the threshold before deployment. As can be seen in **Table 2**, where significant differences are indicated with an asterisk, the majority of the threshold values changed significantly (11 and 9 out of the 17 thresholds for sample 1 and 2, respectively). A decrease in threshold means that the possibility of answering "yes" after deployment was higher than the possibility of a "yes" before deployment, whereas the possibility of answering "yes" was lower after deployment compared to before deployment for those thresholds that increased. According to this method, four items changed significantly in the same direction in both samples: thresholds for "Recurrent distressing dreams of the event," "Restricted range of affect," and "Hypervigilance" decreased, while "Sense of foreshortened future" increased. Only the threshold of three items (i.e., "Acting or feeling as if the event were recurring," "Difficulty falling or staying asleep," and "Difficulty concentrating") did not change significantly in either sample.

The second method was based on chi square differences between either the scalar (method 2A; see **Table 3**) or the loading invariance model (method 2B; see **Table 4**) and 17 models where one combination of thresholds is released or fixed, respectively. Method 2A showed more items with stable thresholds over time, but there was almost no overlap on item level between the two samples. The results of method 2B were similar to the results of method 1, with the only difference that some item thresholds that significantly changed over time according to method 1, did not significantly change according to the *l* value, but only when a *p* value of.05 was used.

In sum, the three methods resulted in different items being problematic and not all items were similarly problematic across the two samples. Looking at the subscales of the PSS (subscales according to the DSM-IV and psychometric studies), each subscale included one or more unstable items. So the main conclusion is that the instrument assessing posttraumatic stress symptoms has way too many non-invariant items to justify latent mean comparison over time.

DISCUSSION

To compare latent mean scores over time, the latent variable should be measurement invariant. However, it might not always be realistic to expect measurement invariance. In the current study we tested whether the underlying construct of

Table 2 | Threshold and threshold difference (threshold after deployment minus threshold before deployment) per item of the Posttraumatic Symptom Scale—Self Report (PSS).

THRESHOLD						
Item	**Sample 1**			**Sample 2**		
	Pre	**Post**	**Diff**	**Pre**	**Post**	**Diff**
1. Recurrent and intrusive distressing recollections of the event	0.221	1.411	1.190*	0.895	0.908	0.049
2. Recurrent distressing dreams of the event	1.440	1.130	−0.310*	1.462	0.990	−0.472*
3. Acting or feeling as if the event were recurring	1.054	1.306	0.252	1.005	0.940	−0.065
4. Intense psychological distress at exposure to cues of event	1.036	1.569	0.533*	1.820	1.060	−0.760*
5. Physiological reactivity on exposure to cues of event	1.258	1.643	0.385*	1.264	1.135	−0.129
6. Avoidance of thoughts, feelings, or conversations associated with event	0.623	1.836	1.213*	1.435	0.762	−0.673*
7. Avoidance of activities, places, or people associated with event	1.036	1.647	0.611*	1.345	1.415	0.070
8. Inability to recall an important aspect of event	0.919	1.356	0.437*	1.191	1.197	0.006
9. Diminished interest or participation in significant activities	0.801	1.021	0.220	1.209	0.668	−0.541*
10. Feeling of detachment or estrangement from others	0.987	1.216	0.229	1.191	0.776	−0.415*
11. Restricted range of affect	1.113	0.890	−0.223*	0.869	0.630	−0.239*
12. Sense of a foreshortened future	1.019	1.359	0.340*	1.017	1.385	0.368*
13. Difficulty falling or staying asleep	0.921	0.830	−0.091	0.820	0.665	−0.155
14. Irritability or outbursts of anger	0.258	0.221	−0.037	0.856	0.273	−0.583*
15. Difficulty concentrating	0.552	0.745	0.193	0.650	0.655	0.005
16. Hypervigilance	0.830	0.330	−0.500*	1.245	−0.166	−0.411*
17. Exaggerated startle response	1.608	0.704	−0.904*	0.694	0.484	−0.210

*$p < 0.05$.

Table 3 | Chi square difference values, p-, and l-values for the scalar model where the model number refers to the item number of which the thresholds between the two time points is estimated unconstrained (all factor loadings and other thresholds are constrained).

Sample 1			**Sample 2**			**l**
Model	χ^2	***p***	**Model**	χ^2	***p***	
M1	77.719	<0.0001*	M16	106.308	<0.0001*	0.00085
M2	17.674	<0.0001*	M12	29.885	<0.0001*	0.00171
M17	54.284	<0.0001*	M15	18.237	<0.0001*	0.00256
M6	48.995	<0.0001*	M6	9.874	0.001*	0.00342
M16	45.051	<0.0001*	M14	9.741	0.001*	0.00427
M11	15.203	0.001*	M4	9.139	0.002*	0.00513
M7	9.590	0.002*	M7	7.512	0.006**	0.00598
M4	7.017	0.008**	M8	6.412	0.011**	0.00684
M14	6.755	0.009**	M9	5.176	0.022**	0.00769
M13	6.493	0.011**	M5	4.235	0.039**	0.00855
M8	5.450	0.020**	M3	3.935	0.047**	0.00940
M5	3.146	0.076***	M13	3.363	0.066***	0.01026
M12	2.296	0.130***	M2	2.789	0.094***	0.01111
M3	1.477	0.224***	M17	1.156	0.282***	0.01197
M10	1.128	0.288***	M10	0.580	0.446***	0.01282
M9	1.088	0.297***	M11	0.485	0.486***	0.01368
M15	0.005	0.942***	M1	0.005	0.941***	0.01453

*significant when $p \leq l$.

**significant when $p \leq 0.05$.

***never significant.

$l = \{0.05/[17*(1+1/2+1/3+1/4+1/5+1/6+1/7+1/8+1/9+1/10+1/11+1/12+1/13+1/14+1/15+1/16+1/17)]\}*c$ where $c = 1,\ldots,17$ to obtain a new alpha value for each new test.

Table 4 | Chi square difference values, p-, and l-values for the loading invariance model where the model number refers to the item number of which the thresholds between the two time points is constrained (all factor loadings are constrained and other thresholds are unconstrained).

Sample 1			**Sample 2**			**l**
Model	χ^2	***p***	**Model**	χ^2	***p***	
M1	92.568	<0.0001*	M16	130.2250	<0.0001*	0.00085
M6	56.579	<0.0001*	M14	27.0260	<0.0001*	0.00171
M16	22.125	<0.0001*	M6	23.6180	<0.0001*	0.00256
M17	35.555	<0.0001*	M9	21.8750	<0.0001*	0.00342
M7	13.277	<0.0001*	M4	21.0990	<0.0001*	0.00427
M4	11.135	0.001*	M10	13.6190	<0.0001*	0.00513
M8	9.798	0.002*	M2	13.4300	0.001*	0.00598
M5	5.807	0.016*	M12	8.4590	0.003*	0.00684
M12	5.232	0.022**	M11	5.9620	0.014**	0.00769
M2	4.890	0.027**	M17	4.3380	0.037**	0.00855
M11	3.969	0.046**	M13	1.8990	0.168***	0.00940
M15	3.960	0.046**	M3	1.2580	0.262***	0.01026
M9	3.890	0.048**	M5	1.0110	0.314***	0.01111
M10	3.497	0.061***	M15	1.0020	0.316***	0.01197
M3	2.777	0.095***	M7	0.2040	0.651***	0.01282
M14	1.132	0.287***	M1	0.1580	0.690***	0.01368
M13	0.607	0.436***	M8	0.0020	0.963***	0.01453

*significant when $p \leq l$.

**significant when $p \leq 0.05$.

***never significant.

$l = \{0.05/[17*(1+1/2+1/3+1/4+1/5+1/6+1/7+1/8+1/9+1/10+1/11+1/12+1/13+1/14+1/15+1/16+1/17)]\}*c$ where $c = 1,\ldots,17$ to obtain a new alpha value for each new test.

a posttraumatic stress questionnaire changed over time by the experience of a traumatic event. This change seems likely, since such a major life experience challenges someone's beliefs about others, the world, and themselves (e.g., Foa and Rothbaum, 1998; Ehlers and Clark, 2000). At the same time, however, assessment of posttraumatic stress before and after a traumatic event is important to study the development op posttraumatic stress disorder after a specific event; that is, already existing symptoms should be taken into account. In the present study, measurement invariance of the posttraumatic symptom scale (PSS; Foa et al., 1993) was tested in two samples of Dutch soldiers who completed the PSS before and after deployment.

According to our first statistical method, results from our test for measurement invariance in Sample 1 showed instability of the thresholds of almost all indicators (the items). Analyses in Sample 2 replicated these findings, but other indicators appeared to be causing the non-invariance. Results were also similar when only those soldiers with or without prior deployment experience were included. Taking both samples into account, only 3 item thresholds showed no significant changes over time. The instability of thresholds was replicated with two other statistical methods, although not all thresholds were similarly problematic across the different methods and the two samples. Since the lack of measurement invariance is due to threshold instability of the majority of the items, it seems reasonable to conclude that the underlying construct of PSS is unstable over time if war-zone related traumatic events occur in between measurements. This finding might also explain the lack of measurement (scalar) invariance found in a study that compared soldiers who had or had not been recently deployed (Mansfield et al., 2010).

From a statistical viewpoint, based on the findings of this study it could be argued that *any* PTSD-related questionnaire is expected to fail the test for measurement invariance. As a result, measurement invariance should never be taken for granted, but should be tested. Moreover, if non-invariance is found, an increase or a decrease of PSS cannot be interpreted in a straightforward way in a prospective longitudinal study in which the PSS is assessed before and after trauma e.g., using, longitudinal models like repeated measure analyses or latent growth (mixture) models. One solution is to treat the pre-trauma assessment as a different construct. Giving the constructs before and after the traumatic event different names can emphasize this: the pre-deployment score could be named "baseline symptoms" (Lommen et al., 2014) and the post-deployment score could be named "PTSD symptoms."

A few points should be taken into consideration with regard to this study. First, although we cross-validated our results in two samples and with different statistical methods, the findings should be replicated in samples from different countries to exclude country specific effects. Also, the results should be replicated in samples with different DSM-classified traumatic events to find out whether the results are specific for military forces or that the results can be generalized to all traumatic events. Moreover, other, more efficient, methods of detecting non-invariant items could be used (de Roover et al., 2014), but at least our conservative method of pairwise testing provides a first step. Future studies may focus on identifying more stable items to construct a questionnaire to use in prospective studies that include measurements before and after trauma exposure. Second, in this study, PTSD was used as a latent construct. The idea that PTSD symptoms are indicators of an underlying latent variable is widespread. According to this view, the PTSD construct denotes a latent variable that functions as the root cause of PTSD symptoms. This presumption has directed psychopathology research for decades, but rests on problematic psychometric premises (Borsboom and Cramer, 2013; McNally et al., in press). Recently, alternative, network approaches have been proposed that conceptualize mental disorders as systems of causally connected symptoms (Borsboom and Cramer, 2013; McNally et al., in press). Future studies might investigate change in PTSD symptoms from a network approach perspective.

RECOMMENDATIONS

Our advice for PTSD researchers who use PTSD as a latent construct in pre-trauma and post-trauma designs is to always test for measurement invariance for measures. Since measurement non-invariance is highly likely to be found if a traumatic event occurred in between two assessments, it is important to investigate the source of the construct instability, and treat the pre and post scores as different construct for each time point in the analysis.

ACKNOWLEDGMENTS

This study was funded by the Netherlands Organization for Scientific Research (NWO) with Vidi grant 452-08-015 and Open Competition grant 400-07-181 awarded to Iris Engelhard. Rens van de Schoot is supported with Veni grant 451-11-008 from NWO.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.01304/abstract

REFERENCES

Akaike, H. (1981). Likelihood of a model and information criteria. *J. Econom.* 16, 3–14. doi: 10.1016/0304-4076(81)90071-3

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders. 4th Edn.,* text revision; *5th Edn.* Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders. 4th Edn.,* text revision; *5th Edn.* Washington, DC: American Psychiatric Association.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238

Berntsen, D., Johannessen, K. B., Thomsen, Y. D., Bertelsen, M., Hoyle, R. H., and Rubin, D. C. (2012). Peace and war: trajectories of posttraumatic stress disorder symptoms before, during, and after military dployment in Afghanistan. *Psychol. Sci.* 23, 1557–1565. doi: 10.1177/0956797612457389

Bonanno, G. A., Mancini, A. D., Horton, J. L., Powell, T. M., LeardMann, C. A., Boyko, E. J., et al. (2012). Trajectories of trauma symptoms and resilience in deployed US military service members: prospective cohort study. *Br. J. Psychiatry*. 200, 317–323. doi: 10.1192/bjp.bp.111.096552

Borsboom, D., and Cramer, A. O. J. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annu. Rev. Clin. Psychol.* 9, 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: The Guilford Press.

Byrne, B. M., Shavelson, R. J., and Muthén, B. O. (1989). Testing for equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

de Roover, K., Timmerman, M. E., De Leersnyder, J., Mesquita, B., and Ceulemans, E. (2014). What's hampering measurement invariance: detecting non-invariant items using clusterwise simultaneous component analysis. *Front. Psychol.* 5:604. doi: 10.3389/fpsyg.2014.00604

Ehlers, A., and Clark, D. M. (2000). A cognitive model of posttraumatic stress disorder. *Behav. Res. Ther.* 38, 319–345. doi: 10.1016/S0005-7967(99)00123-0

Engelhard, I. M., Arntz, A., and van den Hout, M. A. (2007a). Low specificity of symptoms on the post-traumatic stress disorder (PTSD) symptom scale: a comparison of individuals with PTSD, individuals with other anxiety disorders, and individuals without psychopathology. *Br. J. Clin. Psychol.* 46, 449–456. doi: 10.1348/014466507X206883

Engelhard, I. M., de Jong, P. J., van den Hout, M. A., and van Overveld, M. (2009a). Expectancy bias and the persistence of posttraumatic stress. *Behav. Res. Ther.* 47, 887–892. doi: 10.1016/j.brat.2009.06.017

Engelhard, I. M., Olatunji, B. O., and de Jong, P. J. (2011). Disgust and the development of posttraumatic stress among soldiers deployed to Afghanistan. *J. Anxiety Disord.* 25, 58–63. doi: 10.1016/j.janxdis.2010.08.003

Engelhard, I. M., and van den Hout, M. A. (2007). Preexisting neuroticism, subjective stressor severity, and posttraumatic stress in soldiers deployed to Iraq. *Can. J. Psychiatry*. 52, 505–509.

Engelhard, I. M., van den Hout, M. A., and Lommen, M. J. J. (2009b). Individuals high in neuroticism are not more reactive to adverse events. *Pers. Individ. Dif.* 47, 697–700. doi: 10.1016/j.paid.2009.05.031

Engelhard, I. M., van den Hout, M. A., Weerts, J., Arntz, A., Hox, J. J. C. M., and McNally, R. J. (2007b). Deployment-related stress and trauma in Dutch soldiers returning from Iraq: prospective study. *Br. J. Psychiatry*. 191, 140–145. doi: 10.1192/bjp.bp.106.034884

Foa, E. B., Ehlers, A., Clark, D. M., Tolin, D. F., and Orsillo, S. M. (1999). The posttraumatic cognitions inventory (PTCI): development and validation. *Psychol. Assess.* 11, 303–314. doi: 10.1037/1040-3590.11.3.303

Foa, E. B., Riggs, D. S., Dancu, C. V., and Rothbaum, B. O. (1993). Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *J. Trauma. Stress* 6, 459–473. doi: 10.1002/jts.2490060405

Foa, E. B., and Rothbaum, B. O. (1998). *Treating the Trauma of Rape: Cognitive Behavioral Therapy for PTSD*. New York, NY: Guilford Press.

Franz, M. R., Wolf, E. J., MacDonald, H. Z., Marx, B. P., Proctor, S. P., and Vasterling, J. J. (2013). Relationships among predeployment risk factors, warzone-threat appraisal, and postdeployment PTSD symptoms. *J. Trauma. Stress* 26, 1–9. doi: 10.1002/jts.21827

Guenole, N. (2014). Apples, Oranges, and Regression Parameters: consequences of ignoring measurement invariance for path coefficients in structural equation models. *Front. Psychology*. 5:980. doi: 10.3389/fpsyg.2014.00980

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling. 3rd Edn*. New York, NY: The Guilford Press.

Lommen, M. J. J., Engelhard, I. M., Sijbrandij, M., van den Hout, M. A., and Hermans, D. (2013). Pre-trauma individual differences in extinction learning predict posttraumatic stress. *Behav. Res. Ther.* 51, 63–67. doi: 10.1016/j.brat.2012.11.004

Lommen, M. J. J., Engelhard, I. M., van de Schoot, R., and van den Hout, M. A. (2014). Anger: cause or consequence of posttraumatic stress? a prospective study of Dutch soldiers. *J. Trauma. Stress* 27, 200–207. doi: 10.1002/jts.21904

Maguen, S., Litz, B. T., Wang, J. L., and Cook, M. (2004). The stressors and demands of peacekeeping in Kosovo: predictors of mental health response. *Mil. Med.* 169, 198–206.

Mansfield, A. J., Williams, J., Hourani, L. L., and Babeu, L. A. (2010). Measurement invariance of posttraumatic stress disorder symptoms among U.S. military personnel. *J. Trauma. Stress* 23, 91–99. doi: 10.1002/jts.20492

McNally, R. J., Robinaugh, D. J., Wu, G. W. Y., Wang, L., Deserno, M., and Borsboom, D. (in press). Mental disorders as causal systems: a network approach to posttraumatic stress disorder. *Clin. Psychol. Sci*.

Muthén, B. (2014). IRT studies of many groups: the alignment method. *Front. Psychol.* 5:978. doi: 10.3389/fpsyg.2014.00978

Muthén, L. K., and Muthén, B. O. (2010). *Mplus User's Guide. 6th Edn*. Los Angelos, CA: Muthén & Muthén.

Rademaker, A. R., van Zuiden, M., Vermetten, E., and Geuze, E. (2011). Type D personality and the development of PTSD symptoms: a prospective study. *J. Abnorm. Psychol.* 120, 299–307. doi: 10.1037/a0021806

Raykov, T., Marcoulides, G. A., and Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations: a critical view and refocus. *Educ. Psychol. Meas.* 72, 954–974. doi: 10.1177/0013164412441607

Raykov, T., Marcoulides, G. A., and Millsap, R. E. (2013). Factorial invariance in multiple populations: a multiple testing procedure. *Educ. Psychol. Meas.* 73, 713–727. doi: 10.1177/0013164412451978

Rona, R.J., Hooper, R., Jones, M., Iversen, A. C., Hull, L., Murphy, D., et al. (2009). The contribution of prior psychological symptoms and combat exposure to post Iraq deployment mental health in the UK military. *J. Trauma. Stress* 22, 11–19. doi: 10.1002/jts.20383

Schwarz, G. E. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Steenkamp, J. M., and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *J. Consum. Res.* 25, 78–90 doi: 10.1086/209528

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behav. Res.* 25, 173–180. doi: 10.1207/s15327906mbr2502_4

Tucker, L. R., and Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 38, 1–10. doi: 10.1007/BF02291170

Vandenberg, R. J., and Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthén, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

van Zuiden, M., Kavelaars, A., Rademaker, A. R., Vermetten, E., Heijnen, C. J., and Geuze, E. (2011). A prospective study on personality and the cortisol awakening response to predict posttraumatic stress symptoms in response to military deployment. *J. Psychiatr. Res.* 45, 713–719. doi: 10.1016/j.jpsychires.2010.11.013

Vasterling, J. J., Proctor, S. P., Friedman, M. J., Hoge, C. W., Heeren, T., King, L. A., et al. (2010). PTSD symptom increases in Iraq-deployed soldiers: comparison with nondeployed soldiers and associations with baseline symptoms, deployment experiences, and postdeployment stress. *J. Trauma. Stress* 23, 41–51. doi: 10.1002/jts.20487

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 May 2014; accepted: 27 October 2014; published online: 18 November 2014.

*Citation: Lommen MJJ, Van de Schoot R and Engelhard IM (2014) The experience of traumatic events disrupts the measurement invariance of a posttraumatic stress scale. Front. Psychol. **5**:1304. doi: 10.3389/fpsyg.2014.01304*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 12 September 2014
doi: 10.3389/fpsyg.2014.00978

IRT studies of many groups: the alignment method

Bengt Muthén* and Tihomir Asparouhov

Mplus, Los Angeles, CA, USA

Edited by:
Rens Van De Schoot, Utrecht University, Netherlands

Reviewed by:
Eldad Davidov, University of Zurich, Switzerland
Joop J. Hox, Utrecht University, Netherlands
Aleksandra Bujacz, Stockholm University, Sweden

****Correspondence:***
Bengt Muthén, Mplus, 3463 Stoner Ave, Los Angeles, CA 90066, USA
e-mail: bmuthen@statmodel.com

Asparouhov and Muthén (2014) presented a new method for multiple-group confirmatory factor analysis (CFA), referred to as the alignment method. The alignment method can be used to estimate group-specific factor means and variances without requiring exact measurement invariance. A strength of the method is the ability to conveniently estimate models for many groups, such as with comparisons of countries. This paper focuses on IRT applications of the alignment method. An empirical investigation is made of binary knowledge items administered in two separate surveys of a set of countries. A Monte Carlo study is presented that shows how the quality of the alignment can be assessed.

Keywords: factor means invariance testing country comparisons, approximate invariance maximum-likelihood, Bayesian inference, invariance testing, maximum likelihood estimation

1. INTRODUCTION

Asparouhov and Muthén (2014) presented a new method for multiple-group confirmatory factor analysis (CFA), referred to as the alignment method. The alignment method can be used to estimate group-specific factor means and variances without requiring exact measurement invariance. A strength of the method is the ability to conveniently estimate models for many groups, such as with comparisons of countries. The method is a valuable alternative to the currently used multiple-group CFA methods for studying measurement invariance that require multiple manual model adjustments guided by modification indices. Multiple-group CFA is not practical with many groups due to poor model fit of the scalar model and too many large modification indices. In contrast, the alignment method is based on the configural model and essentially automates and greatly simplifies measurement invariance analysis. The method also provides a detailed account of parameter invariance for every model parameter in every group.

This paper focuses on IRT applications of the alignment method. An empirical investigation is made of binary knowledge items administered in two separate surveys of a set of countries. A Monte Carlo study is presented that shows how the quality of the alignment can be assessed. Mplus inputs are provided in the Supplementary Material.

2. MULTIPLE-GROUP IRT

Consider the response to item y expressed by the two-parameter logit model for individual i in group g,

$$P(y_{ig} = 1|\eta_{ig}) = \frac{1}{1 + exp[-a_g(\eta_{ig} - b_g)]}, \quad (1)$$

where $g = 1, \ldots, G$ and G is the number of groups, $i = 1, \ldots, N_g$ where N_g is the number of independent observations in group g, and η_{ig} is a latent variable, $\eta_{ig} \sim N(\alpha_g, \psi_g)$. Using item response theory (IRT) language, a_g is the discrimination parameter and b_g the difficulty parameter. For a recent overview of IRT for psychologists, see e.g., Reise et al. (2013).

Measurement invariance for a_g and b_g (referred to as "item bias" and "DIF" in IRT) has traditionally been concerned with comparing a small number of groups such as with gender or ethnicity using techniques such as likelihood-ratio chi-square testing of one item at a time (see e.g., Thissen et al., 1993). Two common approaches have been discussed (Stark et al., 2006; Lee et al., 2010; Kim and Yoon, 2011):

- Bottom-up: Start with no invariance (configural case), imposing invariance one item at a time.
- Top-down: Start with full invariance (scalar case), freeing invariance one item at a time.

Neither approach is scalable—both are very cumbersome when there are many groups, such as 50 countries ($50 \times 49/2 = 1225$ pairwise comparisons for each item). The correct model may well be far from either of the two starting points, which may lead to the wrong model. Asparouhov and Muthén (2014) proposed a new method referred to as alignment which is suitable for analysis of many groups. The alignment method is based on the idea of starting from the configural model with no invariance and attempting to find as much invariance as possible by letting the factor means and variances vary across groups.

3. THE ALIGNMENT METHOD

Asparouhov and Muthén (2014) considers the model for a continuous item y_{ipg},

$$y_{ipg} = \nu_{pg} + \lambda_{pg}\eta_{ig} + \varepsilon_{ipg}, \quad (2)$$

where $p = 1, \ldots, P$ and P is the number of observed indicator variables, $g = 1, \ldots, G$ and G is the number of groups, $i = 1, \ldots, N_g$ where N_g is the number of independent observations in group g, η_{ig} is a latent variable and we assume that

$\varepsilon_{ipg} \sim N(0, \theta_{pg})$, $\eta_{ig} \sim N(\alpha_g, \psi_g)$. This expression is relevant also for binary outcomes when letting the dependent variable in (2) be a continuous latent response variable y^*_{ipg} underlying the observed binary variable y_{ipg}, where using a threshold parameter τ,

$$y_{ipg} = \begin{cases} 0, & \text{if } y^*_{ipg} \leq \tau_{pg} \\ 1, & \text{if } y^*_{ipg} > \tau_{pg} \end{cases}$$

and the variance of the residual ε_{ipg} is standardized as $\pi^2/3$ in line with the logistic model (with the alternative probit modeling, the residual variance is standardized as one). Using (2), the IRT parameters of (1) are obtained as

$$a_{pg} = \lambda_{pg}, \tag{3}$$

$$b_{pg} = \tau_{pg}/\lambda_{pg}. \tag{4}$$

Asparouhov and Muthén (2014) illustrates the reason for the choice of the term alignment for this new method as in **Figure 1** using continuous items. Consider group-invariant intercepts and loadings for 10 items and two groups with factor means 0 and −1 and factor variances 1 and 2. The configural model of the first step of alignment fixes the factor means and variances to 0 and 1, respectively, in both groups. The plot at the top shows the configural intercept parameters which due to group differences in factor means and variances are not equal across the two groups despite the perfect measurement invariance of the original parameters. The plot at the bottom shows the invariance across groups of the original parameters where the correct factor means and variances have been taken into account. Going from the top to the bottom plot, the intercept parameters have been aligned.

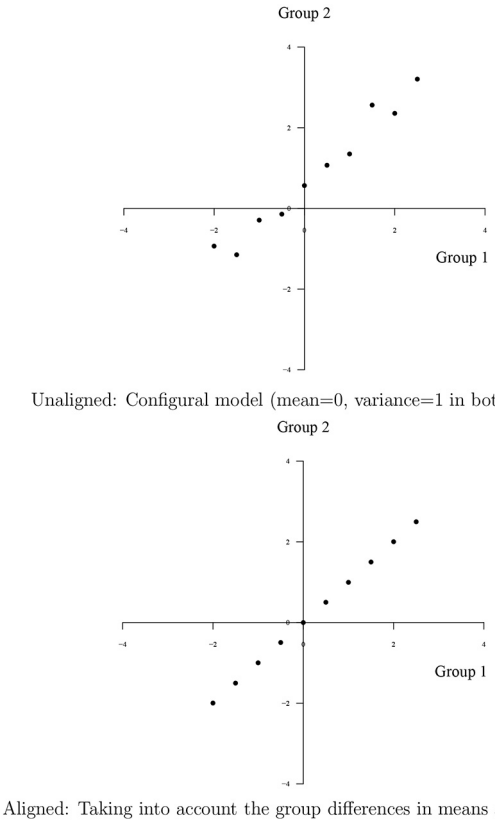


FIGURE 1 | Unaligned and aligned intercept parameters axes correspond to intercept values for the two groups. Unaligned: Configural model (mean = 0, variance = 1 in both groups). Aligned: Taking into account the group differences in means and variances.

3.1. THE ALIGNMENT FITTING FUNCTION

Denote the estimates of the configural model by $\nu_{pg,0}$ and $\lambda_{pg,0}$. Asparouhov and Muthén (2014) show that for every set of parameters α_g and ψ_g there are intercept and loading parameters ν_{pg} and λ_{pg} that yield the same likelihood as the configural model. These parameters can be obtained as follows

$$\lambda_{pg,1} = \frac{\lambda_{pg,0}}{\sqrt{\psi_g}}, \tag{5}$$

$$\nu_{pg,1} = \nu_{pg,0} - \alpha_g \frac{\lambda_{pg,0}}{\sqrt{\psi_g}}. \tag{6}$$

We want to choose α_g and ψ_g so that we minimize the amount of measurement non-invariance. The α_g and ψ_g parameters are, however, not identified in the configural model and are fixed to zero and one, respectively for each group. Adding a simplicity function gives the necessary restrictions to identify the model. The simplicity function minimizes with respect to α_g and ψ_g the total loss/simplicity function F which accumulates the total measurement non-invariance over the items,

$$F = \sum_p \sum_{g_1 < g_2} w_{g_1,g_2} f(\lambda_{pg_1,1} - \lambda_{pg_2,1}) + \sum_p \sum_{g_1 < g_2} w_{g_1,g_2} f(\nu_{pg_1,1} - \nu_{pg_2,1}). \tag{7}$$

The function F implies that for every pair of groups and every intercept and loading parameter we add to the total loss function the difference between the parameters scaled via the component loss function (CLF) f. CLF has been used in EFA analysis, see for example Jennrich (2006) and it is used similarly here. One good choice for the CLF is

$$f(x) = \sqrt{\sqrt{x^2 + \epsilon}}$$

where ϵ is a small number such as 0.0001. Thus, the total loss function F will be minimized at a solution where there are a few large non-invariant measurement parameters and many approximately invariant measurement parameters rather than many medium-sized non-invariant measurement parameters. This is similar to the fact that EFA rotation functions aim for either large or small loadings, but not mid-sized loadings.

The alignment method is carried out using maximum-likelihood estimation of the configural model. In addition to the logit model, probit can also be handled. More than one factor can also be accommodated in which case the alignment is done for each factor. Cross-loadings are not, however, allowed. To handle national surveys, the estimation allows complex survey data with stratification, weights, and clustering, where standard errors are computed using the Huber-White sandwich estimator.

Muthén and Asparouhov (2013a) make a comparison of the alignment method and two-level IRT modeling. In the former approach the groups are viewed as a fixed mode of variation, whereas in the latter approach they are viewed as a random mode of variation. A key advantage of the alignment method is that a specific distributional assumption such as normality of the item parameter distributions across groups is not required. For example, a subset of the groups may show large non-invariance, whereas the remaining groups may show little invariance. Information about which groups contribute to non-invariance is also more readily available with the alignment method.

3.2. ALIGNMENT QUALITY AND DEGREE OF NON-INVARIANCE

In discussing the quality of the alignment results, Asparouhov and Muthén (2014) stated

> "The alignment method will always estimate the simplest model with the largest amount of invariance, but if the assumption of approximate measurement invariance is violated the simplest and most invariant model may not be the true model. For example, if data are generated where a minority of the factor indicators have invariant measurement parameters and the majority of the indicators have the same amount of non-invariance, the alignment method will choose the non-invariant indicators as the invariant ones, singling out the other indicators as non-invariant."

The Asparouhov and Muthén (2014) simulation results show that alignment parameter biases increase with increasing degree of measurement non-invariance, decreasing group sample size, and increasing number of groups. For 60 groups, satisfactory results were obtained with groups sizes of 1000 and at most 20% non-invariant measurement parameters. A key issue is the quality of the ranking of groups based on factor means. Monte Carlo simulations in Muthén and Asparouhov (2013a) focused on the correlation between the population factor means and the estimated alignment factor means computed over groups and averaged over replications. Correlations of at least 0.98 were deemed to produce reliable factor mean rankings. Correlations of this magnitude were seen even in cases with higher than 20% non-invariant measurement parameters. As a rough rule of thumb, a limit of 25% non-invariance may be safe for trustworthy alignment results, while with higher percentages a Monte Carlo simulation study is recommended. Such a study is illustrated below.

4. AN ILLUSTRATION COMPARING COUNTRIES IN TWO CROSS-SECTIONAL SURVEYS

The IEA (International Association for the Evaluation of Educational Achievement) civic knowledge test of 1999 consists of 38 dichotomously scored items. This test, referred to as CIVED, was administered to nearly 90,000 14-year-old students in 28 countries (Torney-Purta et al., 2001; Schultz and Sibberns, 2004). A later survey referred to as ICCS (International Civic and Citizenship Education Study) was carried out in 2009 including 17 link items to make scores comparable to those of 1999 (Schultz et al., 2010). ICCS surveyed over 140,000 eight grade students in 38 countries. 17 countries had comparable national samples and test items and therefore allow comparisons to be made between CIVED achievement and ICCS achievement. Three of these countries had missing data for everyone on at least one of the items at one of the surveys, leaving 14 countries to be compared between the 1999 CIVED and the 2009 ICCS in the current analyses. To further sharpen the comparison, the analyses are restricted to 14-year olds. The IRT alignment analyses to be reported thereby focus on the 17 link items and 29,449 students in 14 countries of CIVED and 10,643 students in 14 countries of ICCS. The 14 countries (country number and country acronym given in parentheses) are: Chile (04; CHL), Colombia (05; COL), Czech Republic (07; CHE), England (09; ENG), Finland (11; FIN), Greece (13; GRC), Italy (16; ITA), Latvia (17; LVA), Norway (19; NOR), Poland (20; POL), Slovak Republic (24; SVK), Slovenia (25; SVN), Sweden (26; SWE), and Switzerland (27; CHE).

Before doing the alignment analysis it is of interest to study measurement invariance using traditional methods, namely comparing the configural, metric, and scalar models (see Muthén and Asparouhov, 2013a). The metric model specifies invariant loadings. The scalar model is of particular interest because it specifies measurement invariance of both thresholds and loadings, a requirement for comparing factor means using traditional methods. **Table 1** shows the results for the 1999 CIVED data, the 2009 ICCS data, and the combined data. It is clear that both the metric and the scalar models are rejected by the likelihood-ratio chi-square tests. Part of the reason for this is that the sample sizes are large so there is considerable power to reject invariance. Although criteria such as difference in global fit indices like CFI or RMSEA (Chen, 2007) or detection of local misspecification (Saris et al., 2009) have been proposed to somewhat mitigate this power issue, they are not available with the maximum-likelihood estimation of binary items considered here.

Whatever step-wise non-invariance search method is applied, a large effort is required to find subsets of items that fulfill scalar invariance sufficiently well in different subsets of the groups. The advantage of the alignment method is that metric and scalar invariance are not required. Instead, factor means are made comparable while minimizing measurement non-invariance.

A 14-group alignment analysis of the 17 items is performed for the 14 countries in each of the two surveys, followed by a 28-group alignment analysis of the two surveys jointly. The joint analysis makes it possible to compare factor means and factor variances not only across countries but also across the two surveys. The survey-specific analyses are used to check that the ordering of countries is not largely affected by considering the two surveys together. It was found that the country ordering was almost exactly the same within studies as in the joint 28-group alignment analysis.

The results of the 28-group joint analysis are shown in **Tables 2**, **3** in factor analysis metric for thresholds and loadings,

Table 1 | Configural, metric, and scalar invariance.

INVARIANCE TESTING - CIVED1999 (14 GROUPS)			
Model	**Number of parameters**	**Loglikelihood**	
Configural	489	−343840.898	
Metric	281	−344830.191	
Scalar	73	−354806.259	
Models compared	**Chi-square**	**Degrees of freedom**	***P*-value**
Metric against configural	1331.149	208	0.0000
Scalar against configural	13535.800	416	0.0000
Scalar against metric	11375.032	208	0.0000
INVARIANCE TESTING - ICSS2009 (14 GROUPS)			
Model	**Number of parameters**	**Loglikelihood**	
Configural	489	−126423.673	
Metric	281	−126779.127	
Scalar	73	−130742.955	
Models compared	**Chi-square**	**Degrees of freedom**	***P*-value**
Metric against Configural	580.862	208	0.0000
Scalar against Configural	7110.001	416	0.0000
Scalar against Metric	6573.006	208	0.0000
INVARIANCE TESTING - CIVED1999 AND ICSS2009 (28 GROUPS)			
Model	**Number of parameters**	**Loglikelihood**	
Configural	979	−493498.177	
Metric	547	−494909.372	
Scalar	115	−509271.808	
Models compared	**Chi-square**	**Degrees of freedom**	***P*-value**
Metric against configural	2083.617	432	0.0000
Scalar against configural	22223.702	864	0.0000
Scalar against metric	19349.849	432	0.0000

Table 2 | Invariance results for aligned threshold parameters for items Y1 to Y17 (numbers in parentheses refer to countries that show significant non-invariance for the parameter).

Y1	**(104)** 105 **(107) (109)** 111 113 116 117 119 120 124 125 **(126)** 127 **(204)** 205 **(207)** 209 **(211)** 213 216 217 219 **(220)** 224 225 226 227
Y2	**(104) (105)** 107 **(109)** 111 **(113) (116)** 117 **(119)** 120 **(124) (125) (126) (127) (204) (205)** 207 **(209)** 211 **(213) (216)** 217 **(219)** 220 224 225 **(226) (227)**
Y3	**(104) (105)** 107 109 111 113 **(116)** 117 119 120 124 125 **(126)** 127 **(204) (205) (207)** 209 211 213 216 **(217)** 219 220 224 225 226 227
Y4	104 **(105)** 107 **(109)** 111 113 **(116) (117)** 119 120 124 125 126 127 204 205 207 209 211 213 216 217 219 220 224 225 226 227
Y5	104 105 107 109 111 113 116 117 **(119)** 120 124 125 **(126)** 127 **(204) (205)** 207 209 **(211) (213)** 216 **(217) (219)** 220 224 225 **(226)** 227
Y6	**(104) (105)** 107 **(109)** 111 **(113) (116)** 117 119 120 **(124)** 125 126 **(127)** 204 205 207 209 211 **(213) (216)** 217 219 220 **(224)** 225 226 227
Y7	**(104) (105)** 107 109 111 113 116 117 119 120 124 125 126 **(127)** 204 205 **(207)** 209 211 213 216 **(217)** 219 220 224 225 226 227
Y8	**(104)** 105 107 109 111 113 116 117 119 **(120)** 124 **(125) (126)** 127 **(204)** 205 **(207)** 209 211 213 216 217 219 **(220) (224)** 225 226 **(227)**
Y9	**(104) (105) (107) (109) (111) (113)** 116 **(117) (119)** 120 **(124)** 125 **(126) (127) (204) (205) (207)** 209 **(211)** 213 216 **(217) (219)** 220 **(224)** 225 **(226)** 227
Y10	104 105 107 **(109) (111) (113)** 116 117 **(119)** 120 124 125 **(126)** 127 204 205 **(207)** 209 **(211)** 213 216 217 219 220 224 225 **(226)** 227
Y11	104 **(105)** 107 109 111 113 116 **(117)** 119 **(120)** 124 125 126 127 204 **(205) (207) (209)** 211 213 216 217 219 220 224 225 **(226)** 227
Y12	**(104) (105) (107) (109) (111)** 113 **(116)** 117 119 **(120)** 124 **(125)** 126 **(127)** 204 205 207 **(209)** 211 213 216 217 219 220 224 225 226 227
Y13	**(104) (105)** 107 **(109)** 111 **(113)** 116 **(117)** 119 120 124 **(125)** 126 127 204 205 **(207)** 209 211 213 216 217 219 220 224 225 226 227
Y14	104 **(105)** 107 **(109)** 111 **(113)** 116 117 **(119)** 120 **(124) (125)** 126 127 204 **(205)** 207 209 211 **(213)** 216 217 219 220 224 225 226 227
Y15	104 105 **(107) (109) (111)** 113 116 **(117) (119)** 120 124 **(125)** 126 **(127)** 204 **(205)** 207 **(209)** 211 213 216 **(217) (219)** 220 224 225 **(226)** 227
Y16	104 105 107 109 111 **(113)** 116 **(117)** 119 120 124 125 **(126) (127) (204)** 205 207 209 211 213 216 **(217)** 219 220 224 225 **(226)** 227
Y17	**(104) (105)** 107 109 111 113 **(116)** 117 119 120 124 **(125)** 126 127 204 205 **(207) (209)** 211 213 216 217 **(219)** 220 **(224)** 225 226 227

The group values correspond to the country coding, where a first digit 1 refers to the CIVED survey, a first digit 2 refers to the ICCS survey, and the next two digits correspond to the country codes given in the text.

respectively. The tables indicate which item parameters are non-invariant in which groups by putting groups in parentheses. It is seen that even after alignment many item parameters remain significantly non-invariant in many of the groups. An interesting feature of alignment is that this does not invalidate the alignment method. Thirty three percent of the thresholds and 11% of the loadings are found non-invariant, averaging to 22% non-invariance. Using the 25% rule of thumb mentioned earlier, this implies trustworthy alignment results. To support this conclusion, Monte Carlo simulations reported in Section 5 based on these parameter estimates show that the factor means are well estimated so that a group comparison can be made.

The results in **Tables 2**, **3** can be augmented by the contributions each item and each group makes to the simplicity function (7). It is of interest to see which items and which groups contribute the most and the least to the non-invariance as quantified by this function. The results can be studied for thresholds and loadings separately or together for an item. It is found that the two least invariant items are items 2 and 9 and the most invariant item is item 4. This largely agrees with

Table 3 | Invariance results for aligned loadings for items Y1 to Y17 (numbers in parentheses refer to countries that show significant non-invariance for the parameter).

Y1	104 105 107 **(109)** 111 113 **(116)** 117 **(119)** 120 124 125 126 127 204 205 207 209 211 213 216 217 219 220 224 225 226 227
Y2	104 **(105)** 107 109 111 113 **(116)** 117 119 120 124 125 126 **(127)** 204 205 207 209 211 213 216 217 219 220 224 225 226 227
Y3	104 105 107 109 111 113 116 117 **(119) (120) (124) (125)** 126 127 204 205 207 209 211 213 216 217 219 220 224 225 226 227
Y4	104 105 107 109 111 113 116 117 119 120 124 125 126 127 204 205 207 209 211 213 216 217 219 220 224 225 226 227
Y5	104 105 107 109 111 113 116 117 **(119)** 120 124 125 **(126)** 127 204 205 207 209 211 213 216 217 219 220 224 225 226 227
Y6	104 105 107 109 111 **(113)** 116 117 **(119) (120)** 124 125 126 127 204 205 207 209 **(211)** 213 216 217 219 220 224 225 226 227
Y7	104 105 107 109 **(111) (113)** 116 117 119 120 124 125 126 127 204 205 207 209 211 **(213) (216)** 217 219 220 224 225 226 227
Y8	104 105 **(107)** 109 **(111)** 113 116 117 **(119)** 120 124 125 126 **(127)** 204 205 **(207)** 209 211 213 216 217 219 220 224 225 226 **(227)**
Y9	104 105 107 109 **(111) (113)** 116 117 119 120 124 **(125)** 126 127 **(204)** 205 **(207)** 209 **(211)** 213 216 217 219 220 224 225 226 227
Y10	104 105 107 109 111 113 116 117 119 120 124 125 126 127 204 **(205)** 207 209 211 213 216 217 219 220 224 225 226 227
Y11	104 105 107 **(109)** 111 113 116 117 119 120 124 125 126 127 204 **(205)** 207 209 211 213 216 217 219 220 224 225 226 227
Y12	**(104)** 105 107 **(109)** 111 113 **(116)** 117 119 120 124 125 126 127 **(204)** 205 207 209 211 213 216 217 219 220 224 225 226 227
Y13	104 105 107 109 111 113 116 117 119 120 124 125 126 127 204 205 207 209 211 213 216 217 219 220 224 225 226 227
Y14	104 105 107 109 111 **(113)** 116 117 119 120 124 125 126 127 204 205 207 209 211 213 216 217 219 220 224 225 **(226)** 227
Y15	104 105 **(107) (109) (111) (113)** 116 117 119 **(120)** 124 125 126 127 204 205 **(207) (209)** 211 213 216 217 219 220 224 225 226 227
Y16	104 105 107 109 111 113 116 **(117)** 119 120 124 **(125) (126)** 127 204 205 207 209 211 213 216 217 219 220 224 225 **(226)** 227
Y17	104 105 107 109 111 113 116 117 119 120 124 125 126 127 204 205 207 209 211 213 216 217 219 220 224 225 226 227

The group values correspond to the country coding, where a first digit 1 refers to the CIVED survey, a first digit 2 refers to the ICCS survey, and the next two digits correspond to the country codes given in the text.

the significance findings in **Tables 2**, **3**. Further inspection of these items is therefore warranted. None of the 28 groups stands out as contributing substantially more to the simplicity function, while three groups stand out as contributing the least to the simplicity function: 225 (Slovenia at the second survey), 213 (Greece at the second survey), and 219 (Norway at the second survey).

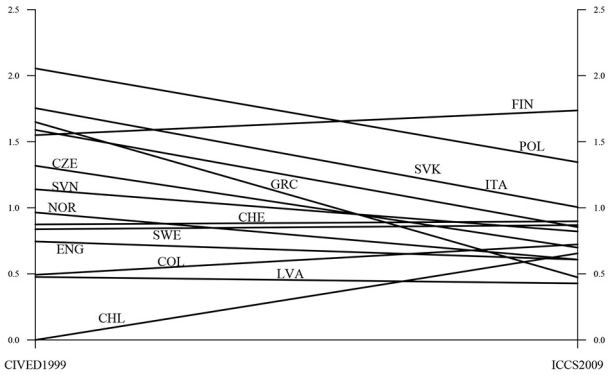


FIGURE 2 | Factor means for CIVED 1999 and ICCS 2009.

The aligned factor means are shown in **Table 4**. The tables also show results of testing for significant factor mean differences between the countries. **Figure 2** gives a graphic representation of factor means at the two surveys. It is seen that a majority of the countries decrease in achievement over the 10 years. Exceptions are Finland, the Czech Republic, Sweden, Columbia, and Chile. The variation in the factor means is also diminished such that fewer countries are at the high end on the factor in 2009 as compared to 1999. It is of interest for test developers to investigate if the causes of these features are partly due to testing artifacts. Such an investigation may include studying differences in item order in the testing booklets, different missing data patterns, and different motivation among the students.

5. MONTE CARLO INVESTIGATION

A useful augmentation of the alignment analysis is to carry out a Monte Carlo simulation study to check how well the factor means are captured. Studies may show a large degree of measurement non-invariance, that is, many measurement parameters show large non-invariance in many groups. The concern may then be that the factor means are not well enough estimated to afford a trustworthy comparison across the groups.

The Monte Carlo study can be done using the same features as in the real-data analysis. The features include the degree of measurement non-invariance, the group-varying factor means and variances, the number of items, the number of groups, and the sample sizes in the groups. Such a Monte Carlo analysis is easily carried out using Mplus. The estimated parameters in the real-data alignment analysis can be saved and used for data generation. A large number of replications (random samples of observations) is used. Summary statistics are provided that include the correlation between the generated and estimated factor means for the countries. A near-perfect correlation is required for the ordering of groups with respect to the factors to be trustworthy. Muthén and Asparouhov (2013a) observed that a correlation of at least 0.98 is needed. For the current 28-group analysis a correlation of 0.996 is observed suggesting excellent alignment despite the non-invariance. The parameter values are also well recovered. Mplus input excerpts for both the real-data and Monte Carlo analyses are shown in the Supplementary Material.

Table 4 | Factor means.

Ranking	Group value	Factor mean	Groups with significantly smaller factor mean
1	120	2.055	113 124 111 220 107 125 216 119 227 127 226 224 126 225 109 205 207 204 209 219 105 117 213 217 104
2	16	1.754	220 107 125 216 119 227 127 226 224 126 225 109 205 207 204 209 219 105 117 213 217 104
3	211	1.737	220 107 125 216 119 227 127 226 224 126 225 109 205 207 204 209 219 105 117 213 217 104
4	113	1.649	220 107 125 216 119 227 127 226 224 126 225 109 205 207 204 209 219 105 117 213 217 104
5	124	1.589	107 125 216 119 227 127 226 224 126 225 109 205 207 204 209 219 105 117 213 217 104
6	111	1.550	107 125 216 119 227 127 226 224 126 225 109 205 207 204 209 219 105 117 213 217 104
7	220	1.345	216 119 227 127 226 224 126 225 109 205 207 204 209 219 105 117 213 217 104
8	107	1.318	216 119 227 127 226 224 126 225 109 205 207 204 209 219 105 117 213 217 104
9	125	1.140	127 226 224 126 225 109 205 207 204 209 219 105 117 213 217 104
10	216	1.005	109 205 207 204 209 219 105 117 213 217 104
11	119	0.965	109 205 207 204 209 219 105 117 213 217 104
12	227	0.898	209 219 105 117 213 217 104
13	127	0.874	209 219 105 117 213 217 104
14	226	0.869	204 209 219 105 117 213 217 104
15	224	0.854	209 105 117 213 217 104
16	126	0.838	209 219 105 117 213 217 104
17	225	0.821	105 117 217 104
18	109	0.745	105 117 217 104
19	205	0.723	217 104
20	207	0.699	117 217 104
21	204	0.655	217 104
22	209	0.608	104
23	219	0.608	104
24	105	0.493	104
25	117	0.477	104
26	213	0.474	104
27	217	0.428	104
28	104	0.000	

The group values correspond to the country coding, where a first digit 1 refers to the CIVED survey, a first digit 2 refers to the ICCS survey, and the next two digits correspond to the country codes given in the text.

6. CONCLUSIONS

The alignment method provides a convenient and powerful method to study IRT modeling in many groups. In recent research 92 groups has proved feasible (Munck et al., 2014). With country comparison it is expected that a large degree of non-invariance is present due to cultural and other country differences. Existing methods are simply not practical for handling such complexity. In the current paper maximum-likelihood estimation was used but Bayesian analysis is also available as discussed in Muthén and Asparouhov (2013a). Bayesian analysis also makes it possible to relax the assumptions of the configural IRT model, for example by allowing certain residual correlations among the items. Bayesian analysis also makes it possible to base the alignment on a model with approximate measurement invariance as discussed in Muthén and Asparouhov (2013b).

Future developments of the alignment method for IRT applications include allowing for different booklets administered to different student groups, adding covariates to the alignment method, and the possibility to create plausible values of the factor scores for secondary analyses. These developments should make IRT alignment an even more valuable addition to the IRT methods arsenal.

ACKNOWLEDGMENT

We thank Ingrid Munck for helpful advice and preparation of the data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00978/abstract

REFERENCES

Asparouhov, T., and Muthén, B. (2014). Multiple-group factor analysis alignment. *Struct. Equ. Modeling* 21, 1–14. doi: 10.1080/10705511.2014.919210

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling* 14, 464–504. doi: 10.1080/10705510701301834

Jennrich, R. (2006). Rotation to simple loadings using component loss functions: the oblique case. *Psychometrika* 71, 173–191. doi: 10.1007/s11336-003-1136-B

Kim, E. S., and Yoon, M. (2011). Testing measurement invariance: a comparison of multiple-group categorical CFA and IRT. *Struct. Equ. Modeling* 18, 212–228. doi: 10.1080/10705511.2011.557337

Lee, J., Little, T. D., and Preacher, K. J. (2010). "Methodological issues in using structural equation models for testing differential item functioning," in *Cross-Cultural Analysis: Methods and Applications*, ed E. Davidov, P. Schmidt, and J. Billiet (New York, NY: Routledge), 55–84.

Munck, I., Barber, C., and Torney-Purta, J. (2014). *Measurement Invariance in International Comparisons of Youth Attitudes Towards Immigrants: the Alignment Method Applied to IEA CIVED 1999 and ICCS 2009.* (in press).

Muthén, B., and Asparouhov, T. (2013a). *New Methods for the Study of Measurement Invariance with Many Groups.* Technical report. Available online at: http://statmodel2.com/download/PolAn.pdf.

Muthén, B., and Asparouhov, T. (2013b). *BSEM Measurement Invariance Analysis. Mplus Web Note 17.* Available online at: http://www.statmodel.com/examples/webnotes/webnote17.pdf.

Reise, S. P., Moore, T. M., Haviland, M. G. G. (2013). "Applying unidimensional item response theory models to psychological data," in *APA Handbook of Testing and Assessment in Psychology, Vol. 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology. APA Handbooks in Psychology*, eds F. Kurt, B. A. Bracken, J. F. Carlson, J. C. Hansen, N. R. Kuncel, S. P. Reise,

et al. (Washington, DC: American Psychological Association, xxix), 101–119. doi: 10.1037/14047-006

Saris, W. E., Satorra, A., and van der Veld W. M. (2009). Testing structural equation models or detection of misspecifications? *Struct. Equ. Modeling* 16, 561–582. doi: 10.1080/10705510903203433

Schultz, W., and Sibberns, H. (2004). *IEA Civic Education Study Technical Report*. Amsterdam: The International Association for the Evaluation of Educational Achievement (IEA).

Schultz, W., Ainley, J., Fraillon, J., Kerr, D., and Losito, B. (2010). *ICCS 2009 International Report: Civic Knowledge, Attitudes, and Engagement Among Lower-Secondary School Students in 38 Countries.* Amsterdam: The International Association for the Evaluation of Educational Achievement (IEA).

Stark, S., Chernyshenko, O. S., and Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *J. Appl. Psychol.* 91, 1292–1306. doi: 10.1037/0021-9010.91.6.1292

Thissen, D., Steinberg, L., and Wainer, H. (1993). "Detection of differential item functioning using the parameters of item response models," in *Differential Item Functioning*, ed P. W. Holland and H. Wainer (Hillsdale, NJ: Lawrence Erlbaum), 67–113.

Torney-Purta, J., Lehman, R., Oswald, H., and Schulz, W. (2001). *Citizenship and Education in Twenty-Eight Countries.* Amsterdam: The International Association for the Evaluation of Educational Achievement (IEA).

Conflict of Interest Statement: The authors are developers of the Mplus software used in the paper. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 April 2014; accepted: 19 August 2014; published online: 12 September 2014.

Citation: Muthén B and Asparouhov T (2014) IRT studies of many groups: the alignment method. Front. Psychol. ***5****:978. doi: 10.3389/fpsyg.2014.00978*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 23 October 2013
doi: 10.3389/fpsyg.2013.00770

Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance

Rens van de Schoot[1,2]*, Anouck Kluytmans[1], Lars Tummers[3,4], Peter Lugtig[1], Joop Hox[1] and Bengt Muthén[5]

[1] Department of Methods and Statistics, Faculty of Social Sciences, Utrecht University, Utrecht, Netherlands
[2] Optentia Research Program, Faculty of Humanities, North-West University, Vanderbijlpark, South Africa
[3] Department of Public Administration, Erasmus University Rotterdam, Rotterdam, Netherlands
[4] Center for the Study of Law & Society, University of California, Berkeley, USA
[5] Graduate School of Education, University of California, Los Angeles, CA, USA

Edited by:
Peter Schmidt, International Labaratory for Socio-Cultural Research HSE Moscow, Russia

Reviewed by:
Oi-Man Kwok, Texas A&M University, USA
Jelte M. Wicherts, Tilburg University, Netherlands

***Correspondence:**
Rens van de Schoot, Department of Methods and Statistics, Utrecht University, Padualaan 14, 3584CH, P.O. Box 80.140, 3508TC Utrecht, Netherlands
e-mail: a.g.j.vandeschoot@uu.nl

Measurement invariance (MI) is a pre-requisite for comparing latent variable scores across groups. The current paper introduces the concept of approximate MI building on the work of Muthén and Asparouhov and their application of Bayesian Structural Equation Modeling (BSEM) in the software M*plus*. They showed that with BSEM exact zeros constraints can be replaced with approximate zeros to allow for minimal steps away from strict MI, still yielding a well-fitting model. This new opportunity enables researchers to make explicit trade-offs between the degree of MI on the one hand, and the degree of model fit on the other. Throughout the paper we discuss the topic of approximate MI, followed by an empirical illustration where the test for MI fails, but where allowing for approximate MI results in a well-fitting model. Using simulated data, we investigate in which situations approximate MI can be applied and when it leads to unbiased results. Both our empirical illustration and the simulation study show approximate MI outperforms full or partial MI In detecting/recovering the true latent mean difference when there are (many) small differences in the intercepts and factor loadings across groups. In the discussion we provide a step-by-step guide in which situation what type of MI is preferred. Our paper provides a first step in the new research area of (partial) approximate MI and shows that it can be a good alternative when strict MI leads to a badly fitting model and when partial MI cannot be applied.

Keywords: measurement invariance, Bayesian structural equation modeling, Mplus, informative/subjective prior, prior variance

INTRODUCTION

If scores on a latent variable are to be compared across groups or time in a meaningful way, the underlying measurement model should be equivalent. Measurement invariance (MI) implies that (for continuous observed variables), conditional on the latent trait scores, the covariances and the intercepts are equal across groups (cf. Mellenbergh, 1989). In other words, the relationships between the latent trait scores and the observed variables do not depend on group membership. Studies of so-called "measurement invariance" have often shown that the underlying constructs are, however, *not* equivalent (e.g., Vandenberg and Lance, 2000; Schmitt and Kuljanin, 2008; Millsap, 2011). The current paper discusses approximate MI as a possible solution to these situations, thereby building on the work of Muthén and Asparouhov (2012b, 2013). Muthén and Asparouhov describe a novel method where, using Bayesian structural equation models (BSEM), exact zero constraints can be replaced with approximate zero constraints based on substantive theories. For example, cross-loadings in confirmatory factor analysis are traditionally constrained to be zero, but using the procedure of Muthén and Asparouhov (2012b) these parameters can be estimated with some, as we call it, "wiggle room" (Muthén and Asparouhov, 2012a), implying that very small cross-loadings are allowed. The novel possibility of approximate zero constraints is an interesting alternative to the use of exact zeros which has proven to be unrealistic at times (see for example van Zuiden et al., 2011). The current paper discusses another area where approximate zeros might have an advantage: when full MI across groups is too strict and small differences in factor loadings or intercepts are allowed to make the model fit well. Possibly differences in use of the response scale are described in Morren et al. (2011).

Muthén and Asparouhov (2013) use the BSEM approach as a way to get the non-invariance information as you would get by Maximum Likelihood (ML) modification indices. They propose a two-step procedure where one first uses BSEM's approximate MI analysis to get modification indices and then free those non-invariant parameters in a regular Bayes run as a final, second step. BSEM modification indices are helpful, for example, when having categorical items where no ML modification indices exist, or with a large number of groups. This is often the case in the

context of large scale international studies. In the current paper we focus on the benefits or dangers when applying approximate invariance when it is actually applied in a CFA model. As we will show with both an empirical illustration and a simulation study, approximate MI enables the researchers to make explicit trade-offs between the degree of MI on the one hand, and the degree of model fit on the other. However, as our simulation results demonstrate, some bias in the estimated parameters occurs due to the alignment issue (see also Muthén and Asparouhov, 2013), which can be corrected using a method available in Mplus v7.1 (Asparouhov and Muthén, 2013).

In what follows we first illustrate issues with applying MI, followed by an introduction of approximate MI. Thereafter, we provide an empirical illustration where the test for strict MI fails, but where approximate MI results in a well-fitting model. Then, with a simulation study, we investigate whether approximate MI can lead to unbiased estimates for differences in latent scores across groups. Thereafter, we introduce the correction method and show its influence on the parameters in our simulation study. We conclude with a discussion and practical recommendations for scholars who aim to meaningfully compare scores on latent variables. Note that the application of approximate MI in the current paper is limited to situations with a small number of groups, continuous variables, and "almost" invariant models. For a more general approach see Muthén and Asparouhov (2013).

THE ISSUE OF APPROXIMATE MEASUREMENT INVARIANCE: SCYLLA OR CHARYBDIS

Questionnaires are often used to assess latent constructs, such as human attitudes and behavior, with the goal to compare groups. For such a comparison to be valid MI should apply, see (Millsap, 2011) or Vandenberg and Lance (2000) for a comprehensive overview on possible methods testing MI. That is, a questionnaire should measure identical constructs with the same factor structure across different groups. Stated differently, factor loadings, intercepts, and residual variances should be identical to get the label "full measurement invariance." If one wants to compare latent means the intercepts are of major importance and therefore, we focus on the intercepts.

Van de Schoot et al. (2012) stated that "When MI does not hold, groups or subjects [. . .] respond differently to the items and as a consequence factor means cannot reasonably be compared" (p. 487). This statement refers to a potential bias in the latent mean comparison when full MI is assumed, but not supported by the data, or when MI is not assumed and the latent means are (incorrectly) compared. In order to meaningfully compare latent means across groups, at least the factor loadings and intercepts should be equal; this is the situation of scalar invariance (Vandenberg and Lance, 2000). Henceforth, when (full) MI is used we refer to scalar invariance. After testing for scalar MI it might be that such a model does not fit the data. What to do in such a situation? One solution is to allow for partial MI. Steenkamp and Baumgartner (1998) suggested that as long as at least two of the factor loadings and intercepts are constrained to be equal across groups or time, the difference in the latent mean between the groups is unbiased (see also Steinmetz, 2013). However, this procedure has been debated much (Vandenberg, 2002), for example how to choose the reference category (Rensvold and Cheung, 2001). At least partial invariance for the factor loadings before one can proceed to test invariance of the intercepts (Steenkamp and Baumgartner, 1998). This paper focuses on comparison of latent means, so we present approximate MI in the context of the intercepts.

To sum up, if MI is used to either see if measurement instruments are equivalent across populations, or to compare the latent means to each other, possible outcomes of MI are:

(1) (full or) scalar MI, where all intercepts are constrained to be equal across groups.
(2) partial MI, where some of the intercepts between groups are allowed to be freely estimated, while others are held constant (see e.g., Steenkamp and Baumgartner, 1998; Steinmetz, 2013); or
(3) No invariance, where all intercepts between groups are freely estimated, because such a model fits the data best. Consequently, the questionnaire cannot be used for comparing groups.

In the current paper we add a fourth option, initiated by Muthén and Asparouhov (2012b, 2013) and introduced in more detail below:

(4) Approximate MI, a Bayesian solution allowing for some wiggle room for the intercept differences between groups, where the wiggle room is determined by the degree of precision of the prior.

Metaphorically speaking, in testing for MI one has to choose between Scylla and Charybdis, two mythical Greek sea monsters[1]. In the current paper we apply this metaphor to the procedure of testing for MI. On the one hand, there is the six-headed sea monster Scylla, who metaphorically represents imposing full MI on the model with as a result that the model fit indices indicate a bad fit to the data. On the other hand, however, we could fall victim to Charybdis if we release the constraints. By not imposing MI, our model will fit the data, but it will be impossible to compare groups. This paper illustrates the third option, using approximate MI, which could turn out to be the way to escape both threats.

Consider a CFA model with two groups, see **Figure 1**. Suppose the difference between the intercepts of item 1 is 0.10. Now, we impose MI on this model, by constraining the two intercepts to be equal. As a result, the difference between both will be exactly zero, that is, we are imposing a difference of zero on the parameter estimates for the intercepts. In **Figure 2** the likelihood function (which is a function of the distribution of the data) is shown for

[1]The two monsters occur in an episode of the adventures of Odysseus; their location is believed to have been at the Strait of Messina between Sicily and the Italian mainland. Scylla, a six-headed sea monster, lived on one end of this strait, while on the other Charybdis resided, causing huge whirlpools. The two monsters were living so close to each other that they created an inescapable threat. Sailors who avoided Charybdis were doomed to meet Scylla and vice versa; it seemed almost impossible to pass the sea strait without being confronted with either of the two mythical monsters.

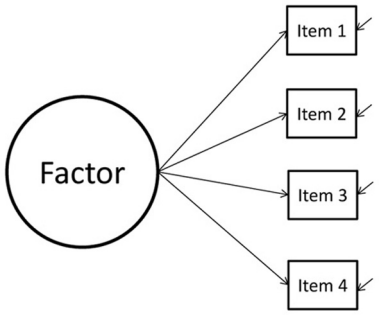


FIGURE 1 | A hypothetical model.

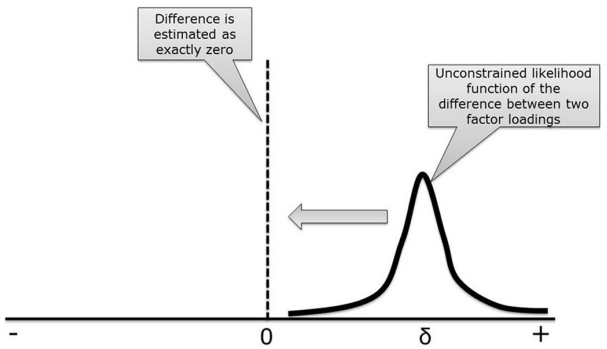


FIGURE 2 | The influence of applying MI while the difference between factor loading is clearly not zero.

the difference between both intercepts, which is denoted by δ. In this case there is a small difference in the intercepts between both groups. When applying MI, the difference is forced to be zero ($\delta = 0$). By doing this, we have established MI and we are allowed to compare the latent factor means between the two groups. However, the estimated intercepts no longer resembles their unconstrained counterparts. Stated differently, δ is forced to be zero, whereas in the data $\delta > 0$. The discrepancy between δ in our model and δ in the data will probably result in poor model fit. A bad model fit means we have to reject our model and cannot interpret our model parameters.

Meanwhile, on the other side of the narrow channel between Italy and Sicily, Charybdis lurks, forced to live in a cave beneath the sea causing whirlpools. If we would analyze our hypothetical model without any constraints on the intercepts the model will fit the data. As a consequence, however, we are lost in the whirlpools caused by the furious Charybdis, because we can no longer compare the latent means due to different intercepts across the groups.

There we are, trapped between Scylla and Charybdis, and are forced to choose between either a model with MI and a terrible fit to the data, or a well-fitting model that we cannot use for comparing the latent means across groups. However, just like Odysseus, we believe we can pass in safety through the narrow channel. One passage may be provided by imposing partial MI allowing for one or two differences. Partial MI seems attractive when relatively large differences ($\delta >> 0$) exist for one or only a few items. However, when differences are small and occur for multiple items in a factor analysis, partial MI is not able to provide a safe passage and a*pproximate* MI offers an attractive alternative. With approximate MI, instead of forcing intercepts to be exactly equal across groups, see **Figure 2**, a substantive prior distribution is used to bring the parameters close to one another while allowing for some wiggle room. Such a model falls in between full and no MI, which could mean that we can still compare the means (as MI holds approximately) while the model also fits well, allowing an escape from Scylla and Charybdis. But how does this work?

USING BAYESIAN PRIORS ON INTERCEPT DIFFERENCES

To estimate a model with approximate MI we need Bayesian statistics, which has been discussed in many papers and textbooks (see, among others, Kruschke et al., 2012; Van de Schoot et al., 2013). There are three essential ingredients underlying Bayesian statistics. The first ingredient is prior distributions which represent background knowledge about the parameters of a model; for example that the difference between two intercepts is close to zero. Second, there is the likelihood function of the data containing the information about the parameters from the data. Thirdly, both prior and likelihood are summarized by the so-called posterior distribution, which is a compromise of the prior knowledge and the likelihood function. Stated otherwise, the posterior distribution contains one's updated knowledge balancing prior knowledge with observed data.

The crucial ingredient of Bayesian statistics is the specification of the prior distribution. In **Figure 3**, four different priors are specified and combined with the likelihood function of the difference between two intercepts which is denoted by δ. When combining prior and likelihood, the posterior difference score is obtained, denoted by δ'. **Figure 3A** displays a flat uninformative prior for the difference between the two intercepts. Because such a prior does not contain any information, the posterior estimate for the difference will not be influenced and the results are similar to a model without MI, that is $\delta = \delta'$. If, for example, a normal prior distribution is used, see **Figure 3B**, the posterior estimate for the difference, δ', will be slightly pulled toward the mean of the prior, in this case zero. If we decrease the prior variance, see **Figure 3C**, the posterior difference comes closer to zero. If the prior variance is very small, the posterior difference will approximate zero, $\delta' \sim 0$, and we establish approximate MI allowing for some wiggle room. To get back to our metaphor: if a small difference between the intercepts is allowed, we can escape Charybdis because the difference between intercepts is smaller than in the unconstrained model, **Figure 3A**. We also escape Scylla, because a model with some wiggle room is less restrictive than full MI and will therefore, still fit the data acceptably well, **Figure 2**. In conclusion, approximate MI finds a compromise between zero and no constraints, through which both model fit and latent mean comparison can be established.

Approximate MI is expected to be especially useful when there are many small deviations from strict MI (De Boeck, 2008; Muthén and Asparouhov, 2013). In the current paper we focus on studying the differences between strict, partial and approximate MI in a set of populations. In the current paper we assume

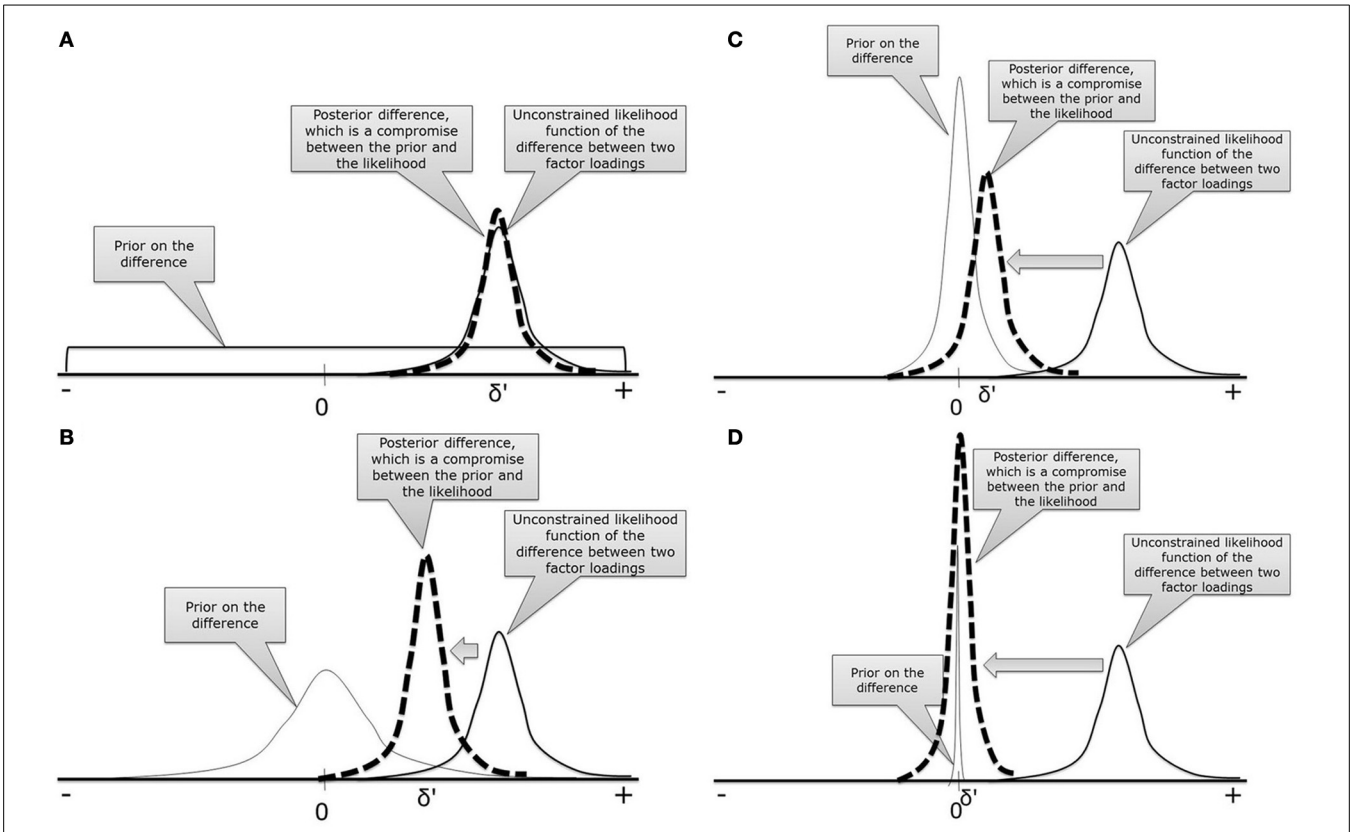


FIGURE 3 | Four different prior distributions to demonstrate the influence of the prior on the posterior parameter estimates. (A) Uninformative prior; **(B)** Wide normal prior; **(C)** Narrow normal prior; **(D)** Highly peaked prior.

that the main goal of applying MI is to compare latent means and, therefore, focus on the potential bias in the latent mean comparison when different degrees of MI are applied. There are two indicators to keep in mind: (1) model fit and (2) a small enough difference between either factor loadings or intercepts.

EMPIRICAL ILLUSTRATION

INTRODUCTION

The empirical illustration looks at the experiences of psychologists (group 1) and psychiatrists (group 2) with a new policy in Dutch mental healthcare: Diagnosis Related Groups (DRGs; Tummers et al., 2012). Diagnosis Related Groups were introduced in January 2008 and were part of a process to convert the Dutch healthcare system into one based on a regulated market. The DRG policy differs significantly from the prior method in which each medical action resulted in a financial claim, a so-called fee-for-service system. Before 2008, the number of sessions a professional had with a patient related directly to the amount of money claimed from the health insurer. According to some standpoints, this could lead to inefficiencies (Busse et al., 2011). The DRG policy changed the situation by stipulating a standard rate for each disorder. For instance, for a mild depression, the mental healthcare professional gets a standard rate for treating the patient (direct and indirect time) between 250–800 min.

Psychologists and psychiatrists had to implement these DRGs, and we will investigate their willingness to do so. This is important, as many of them opposed the DRG policy, set up websites agitating against it, or even in a few cases quit their jobs (Palm et al., 2008). The following quote of a healthcare professional [cited in Tummers (2012): 516], illustrates their point of view:

> *"We experience the DRG policy as a disaster. I concentrate as much as possible on treating my own patients, in order to derive some satisfaction from my work."*

Furthermore, psychiatrists were far more resistant than psychologists. One of the reasons was that especially psychiatrists considered the DRGs as a threat to their autonomy (Smullen, 2013). It is important to analyze the difference between the two groups, in order to provide guidance to policy makers in their attempts to adapt the policy and increase the satisfaction of professional health workers. We would expect minor violations of MI given that the both groups of professionals were expected to be quite negative about the specific policy and also have slightly different attributes to the concepts used in the questionnaires because of their professional training and working environment (see for instance Palm et al., 2008; Neukrug, 2011; Smullen, 2013).

METHODS

The sampling frame consisted of 5199 professionals, all members of the two main nationwide mental healthcare associations: the

Dutch Association of Psychologists (NIP) and the Netherlands Association for Psychiatry (NVvP), who would, in principle, all of them be required to work with the DRG policy. Using an email and two reminders, 1307 questionnaires were returned; a response rate of 25% with 1074 valid cases for the specific scale we used. Despite the select sample the demographical composition of the respondent group was representative for the Dutch population of mental healthcare professionals (Palm et al., 2008).

Willingness to implement the DRG policy was measured using a validated four-item scale developed by Metselaar (1997), which is based on the notion of "intention to act" in the theory of planned behavior (Ajzen, 1991). The items use five-point Likert-scale response categories (strongly disagree, disagree, neutral, agree, and strongly agree). The items use templates in which one can specify the change being assessed, for example, the item "I intend to make time to implement the change" was changed into "I intend to make time to implement the DRG-policy." All item descriptions, its means, variances, and correlations are included in **Table 1** and the data and all syntax files are available on the website of the first author.

RESULTS

If we want to compare psychologists and psychiatrists on the willingness to implement DRGs, we could simply compare the mean scores based on the four items. It appeared that, using a *T*-test in SPSS, psychiatrists ($M = 2.23$; $SD = 0.81$; $n = 504$) indeed scored significantly lower compared to psychologists ($M = 2.46$; $SD = 0.76$; $n = 570$; $M_{dif} = 0.23$; $t = 4.83$; $p < 0.001$). However, by using the mean score we assume that each item reflects the underlying construct in the same way and, even more importantly, that there is no measurement bias (Steinmetz, 2013). To accommodate these unwanted side-effects we conducted a series of confirmatory factor analyses (CFA) using the software M*plus* v7 (Muthén and Muthén, 1998-2012). The data and all syntax files are available as supplementary materials.

In the first model, a 2-group configural model, because of the (slightly) non-normal distributed items estimated with ML estimator with robust standard errors (i.e., MLR), we allowed the factor loadings and intercepts to vary across groups resulted in a well-fitting model ($\chi^2 = 12.982$; $df = 4$; $p = 0.011$; RMSEA = 0.065; CFI = 0.992; TLI = 0.976) with standardized factor loadings ranging between 0.56–0.87. We tested for MI using the new option in Mplus v7.11 `ANALYSIS: MODEL = CONFIGURAL METRIC SCALAR`. A model forcing scalar MI, i.e., factor loadings and intercepts were constrained across groups, appeared to fit the data well ($\chi^2 = 32.032$; $df = 10$; $p < 0.001$; RMSEA = 0.064; CFI = 0.980; TLI = 0.976), but not better compared to the configural model ($\Delta\chi^2 = 19.479$; $\Delta df = 6$; $p = 0.003$). Also the metric model, where only the factor loadings were held equal across groups, fitted the data ($\chi^2 = 18.605$; $df = 7$; $p = 0.009$; RMSEA = 0.056; CFI = 0.990; TLI = 0.982) and not any worse compared to the configural model ($\Delta\chi^2 = 5.019$; $\Delta df = 3$; $p = 0.170$). We also ran a comparison between the scalar and metric model and it appeared that the scalar model fits the data worse compared to the metric model ($\Delta\chi^2 = 13.988$; $\Delta df = 3$; $p = 0.003$). According to most fit indices (e.g., χ^2 not significantly worse than the configural model, but significantly better than the scalar model) the best model appeared to be the metric model where the factor loadings are constrained while the intercepts are allowed to differ across groups.

A solution offered by, for example Byrne et al. (1989; see also Steenkamp and Baumgartner, 1998), is to apply partial MI. To establish partial invariance, one studies the size of the unconstrained loadings and/or intercepts, and constrains all loadings and intercepts except for the one loading/intercept with the largest unstandardized difference, which is released. It appeared that psychiatrists have lower intercepts than the psychologists, with the differences being 0.193, 0.235, 0.167, and 0.324, respectively. We applied partial MI, that is, constraining the intercepts of items 1 and 3 while releasing the constraints on intercepts 2 and 4 ($\chi^2 = 20.271$; $df = 8$; $p = 0.009$; RMSEA = 0.053; CFI = 0.989; TLI = 0.983). Using the procedure described on the website of Mplus to compute MLR chi-square difference testing, it appeared that the partial model did not result in a better fit compared to the metric model ($\Delta\chi^2 = 1.502$; $\Delta df = 1$; $p = 0.203$), but better compared to the scalar model ($\Delta\chi^2 = 12.313$; $\Delta df = 2$; $p = 0.002$).

We re-analyzed the two models, constrained and unconstrained intercepts, using the ML and Bayesian estimator using the default prior settings [i.e., normal prior distributions for the intercepts and factor loadings with a prior mean of zero and a prior variance of 10^{10}, and an inverse gamma distribution for the (residual) variance terms with hyperparameters -1 and zero], but with a stricter cut-off value for convergence to reduce any bias caused by precision [i.e., `Chains = 8`, `Bconvergence = 0.01` and `Biterations(20000)`]. **Table 2** shows the results for the intercepts, the difference between the intercepts, and the Bayesian model fit information. These

Table 1 | Correlation matrix for Psychologists ($n = 570$) and Psychiatrists ($n = 504$) with the means (variances) on the diagonal.

	1	2	3	4
1. I intend to try to convince employees of the benefits the DRG-policy	2.023 (0.727)/ 1.831 (0.730)			
2. I intend to put effort into achieving the goals of the DRG-policy	0.589/0.549	2.651 (1.040)/ 2.414 (1.137)		
3. I intend to reduce resistance among employees regarding the DRG-policy	0.727/0.737	0.616/0.599	2.353 (0.763)/ 2.186 (0.950)	
4. I intend to make time to implement the DRG-policy	0.451/0.470	0.442/0.492	0.483/0.514	2.795 (0.939)/ 2.472 (1.091)

Table 2 | The results for the intercepts of the latent variable *Willingness to Implement DRGs*.

		Model A		Model B		Model C		Model D		Model E		Model F		Model G	
		Measurement invariance		No constraints on the intercepts		Approximate MI $\sigma^2 = 0.50$		Approximate MI $\sigma^2 = 0.05$		Approximate MI $\sigma^2 = 0.01$		Approximate MI $\sigma^2 = 0.005$		Approximate MI $\sigma^2 = 0.0005$	
		ν (*SE*)	95% *CI*	ν (*SE*)	95% *CI*	ν (*SE*)	95% *CI*	ν (*SE*)	95% *CI*	ν (*SE*)	95% *CI*	ν (*SE*)	95% *CI*	ν (*SE*)	95% *CI*
Intercepts group = psychologists	Item 1	2.022 (0.032)	1.961–2.085	2.022 (0.035)	1.955–2.091	2.020 (0.034)	1.954–2.088	2.006 (0.034)	1.943–2.072	1.979 (0.034)	1.957–2.090	1.961 (0.030)	1.904–2.021	1.935 (0.027)	1.885–1.990
	Item 2	2.634 (0.037)	2.563–2.709	2.650 (0.042)	2.569–2.733	2.647 (0.042)	2.565–2.731	2.631 (0.041)	2.550–2.712	2.597 (0.041)	2.569–2.729	2.577 (0.037)	2.506–2.649	2.545 (0.033)	2.483–2.610
	Item 3	2.372 (0.034)	2.308–2.440	2.352 (0.036)	2.281–2.425	2.349 (0.037)	2.278–2.420	2.334 (0.036)	2.264–2.402	2.305 (0.035)	2.285–2.424	2.286 (0.032)	2.224–2.349	2.269 (0.029)	2.212–2.723
	Item 4	2.724 (0.035)	2.657 (2.792)	2.795 (0.041)	2.713–2.876	2.792 (0.041)	2.713–2.868	2.775 (0.040)	2.697–2.851	2.739 (0.040)	2.704–2.859	2.716 (0.036)	2.642–2.783	2.660 (0.032)	2.596–2.723
Intercepts group = psychiatrists	Item 1	2.022 (0.032)	1.961–2.085	1.830 (0.039)	1.757–1.908	1.831 (0.038)	1.758–1.905	1.847 (0.037)	1.771–1.919	1.881 (0.062)	1.917–2.162	1.896 (0.032)	1.836–1.961	1.925 (0.027)	1.873–1.978
	Item 2	2.634 (0.037)	2.563–2.709	2.413 (0.048)	2.323–2.508	2.415 (0.048)	2.323–2.509	2.434 (0.046)	2.346–2.526	2.477 (0.066)	2.503–2.765	2.496 (0.040)	2.420–2.578	2.533 (0.034)	2.468–2.600
	Item 3	2.372 (0.034)	2.308–2.440	2.185 (0.043)	2.103–2.270	2.186 (0.043)	2.103–2.269	2.204 (0.041)	2.124–2.285	2.241 (0.069)	2.287–2.562	2.257 (0.036)	2.188–2.329	2.275 (0.030)	2.217–2.336
	Item 4	2.724 (0.035)	2.657 (2.792)	2.472 (0.046)	2.383–2.562	2.472 (0.046)	2.382–2.563	2.492 (0.045)	2.404–2.581	2.539 (0.058)	2.549–2.777	2.564 (0.039)	2.489–2.643	2.629 (0.033)	2.566–2.695
Difference in intercept	Item 1	0		0.192		0.189		0.159		0.098		0.065		0.010	
	Item 2	0		0.237		0.232		0.197		0.120		0.081		0.012	
	Item 3	0		0.167		0.163		0.130		0.064		0.029		−0.006	
	Item 4	0		0.323		0.320		0.283		0.200		0.152		0.031	
Model fit	95% CI for the difference between the observed and the replicated χ^2	5.128 44.154		−4.164 34.566		−5.516–40.199		−4.369–38.364		−5.543–39.921		3.573–48.979		18.248–60.600	
	Posterior predictive *p*-value	0.008		0.067		0.057		0.062		0.031		0.012		0.000	

results show that a model with strict MI assumed does not fit the data. This is shown by the fact that (1) the posterior predictive *p*-value is significant, and (2) the 95% CI of the replicated Chi Square values does not include zero. Hence, the model without MI does fit the data, but we are not allowed to compare the latent means between psychiatrists and psychologist.

The new option is to use approximate MI. Using Bayesian statistics parameters can be restricted by specifying a prior distribution. We would like the difference between the intercepts to approximate zero, but to allow for some wiggle room (prior variance) to maintain a fitting model. That is, the difference in an intercept between the two groups is allowed to exist, but is restricted to be very small, which is established by specifying a specific prior distribution of that difference. We used the new `DIFF` option available in M*plus* v7 within the `MODEL PRIOR` part of the syntax where subjective priors can be specified. The full syntax is shown in the Appendix A, but the most important part is:

```
MODEL:[Veran1- Veran4] (nu#_1 - nu#_4);

MODEL PRIOR: DO(1,4) DIFF(nu1_#-nu2_#)~N(0,0.50);
```

where `(nu#_1 - nu#_4)` defines labels for the four intercepts. Because we used #,the labels are automatically specified for both groups separately. The `DO(1,4)` option is a loop statement telling M*plus* to apply the function which comes after the `DO` statement for items 1 through 4: #=1 to #=4. The `DIFF` statement refers to the difference between the first intercept of the psychiatrists, for example `nu1_1`, and the same intercept for psychologists, for example `nu1_2`. Because we used the `DO` option this is automatically repeated for all four intercepts. Furthermore, `~N(0, 0.50)` indicates the intercept differences between groups to be normally distributed (N) with mean 0 and prior variance of 0.5 for all pairs of items. Note that we parametrized the model by forcing both latent means to zero and the variance to one.

The results for this specific model are shown in **Table 2** in the column labeled Model C. We varied the prior variance by using $\sigma^2 = 0.05$ (Model D), $\sigma^2 = 0.01$ (Model E), $\sigma^2 = 0.005$ (Model F), and $\sigma^2 = 0.0005$ (Model G).In Model C, with a large prior variance, the difference between intercepts appeared not to be smaller compared to the unconstrained Model B. In Model D, however, the influence of the prior specification can be observed: the difference between intercepts becomes smaller. In Model E the intercepts are even closer, in Model F they are very close and in Model G they are almost similar. However, the latter two models do not fit the data very well; i.e., the 95% CI for the difference between the observed and the replicated χ^2 does not include zero and the *ppp*-value (i.e., posterior predictive *p*-value) is < 0.01. In sum, allowing for a prior variance of 0.01 between the intercepts, as is the case in Model E, resulted in an acceptable model fit. Also, the confidence interval of $\Delta\chi^2$ does include zero. However, the posterior predictive *p*-value is significant, and preferably should be closer to 0.50. A larger reduction, which would be a model closer to scalar invariance, did not fit the data.

To summarize, we have established MI using the newly available approximate MI method. Now, we can finally conclude that psychiatrists score significantly lower on the willingness to implement DRGs than psychologists. The mean difference equals 0.21 ($p < 0.001$), which would indeed be somewhat different had we used full MI ($M_{dif} = 0.19$) or an unconstrained model ($M_{dif} = 0.14$).

However, little is known about the bias of parameters as a result of approximate MI. Therefore, in the next section we will conduct a simulation study to find out if we are truly allowed to interpret the mean difference of the latent mean between groups if we apply approximate MI.

SIMULATION STUDY

METHOD

To investigate the possible bias in the comparison of latent means as a result of applying the approximate MI model we performed a simulation study. Seven populations were specified from which we obtained 1000 datasets each. The difference in intercepts between both groups varied across these seven populations, see **Table 3**. All other parameters were kept constant across populations; see Appendix B for the syntax and model specifications. Most importantly, the mean of the latent variable in group 1 was set to 0 and in group 2 to 0.5. Both latent factors were specified to have a population variance of 1. All items are standardized

Table 3 | Population values for the intercepts.

	Intercepts							
	Item 1		Item 2		Item 3		Item 4	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
Population 1	0	0	0	0	0	0	0	0
Population 2	0	0	0	0	−0.01	0.01	−0.01	0.01
Population 3	0	0	0	0	−0.1	0.1	−0.1	0.1
Population 4	0	0	0	0	−0.5	0.5	−0.5	0.5
Population 5	−0.01	0.01	−0.01	0.01	−0.01	0.01	−0.01	0.01
Population 6	−0.1	0.1	−0.1	0.1	−0.1	0.1	−0.1	0.1
Population 7	−0.5	0.5	−0.5	0.5	−0.5	0.5	−0.5	0.5

making the latent mean difference between the two groups of 0.5 a medium effect size (Cohen, 1992). The sample size per group was specified as being 500.

The seven populations described in **Table 3** were confronted with a set of models:

- Model 1: *scalar* MI is applied to the intercepts *and* factor loadings. Results were obtained with `ESTIMATOR = ML` and with `ESTIMATOR = BAYES`. For the latter we used `BCONVERGENCE = 0.01, BITERATIONS = (5000)`, and the default priors [i.e., normal prior distributions for the intercepts and factor loadings with a prior mean of zero and a prior variance of 10^{10}, and an inverse gamma distribution for the (residual) variance terms with hyperparameters −1 and zero].
- Model 2: *partial* MI is applied to those intercepts that are not similar in the population. For population 1 no partial MI can be applied, since all intercepts are similar in the population, for populations 2–4 partial MI is applied to the intercepts of items 3 and 4, and for populations 5–7 partial MI is applied to all intercepts. Note that the factor loadings are held equal across groups. Results were obtained with `ESTIMATOR = ML` and with `ESTIMATOR = BAYES`. For the latter, we used `BCONVERGENCE = 0.01, BITERATIONS = (5000)`, and the default priors.
- Model 3: *approximate* MI is applied only to the intercepts. We varied the prior variance: $\sigma^2 = 0.5$ (Model 3a), $\sigma^2 = 0.05$ (Model 3b), $\sigma^2 = 0.01$ (Model 3c), and $\sigma^2 = 0.005$ (Model 3d). For all other parameters we used the default prior settings.
- Model 4: *partial* approximate MI, where wiggle room is applied only to those intercepts that are not equal in the population; populations 2–4. We varied the amount of prior variance: $\sigma^2 = 0.5$ (Model 4a), $\sigma^2 = 0.05$ (Model 4b), $\sigma^2 = 0.01$ (Model 4c), and $\sigma^2 = 0.005$ (Model 4d).

The simulated differences in intercepts may cause an alignment issue, i.e., a biased estimate of the latent mean difference across groups, which will be discussed in more details in the next section. Because researchers usually wish to compare latent means across groups, we focus on whether or not the estimated difference in latent means is biased. We focused on four outcome criteria that might indicate the degree to which the mean difference is biased:

(1) the empricial standard deviation of the 1000 estimated mean differences, which should be <0.10.
(2) the relative mean bias defined as $((M - 0.5)/0.5)*100$, where M is the average mean obtained from the simulation study. We used a cut-off value of <10% as a criterion, as suggested by Hoogland and Boomsma (1998) for "reasonable" accuracy.
(3) The proportion of replications with a *ppp*-value smaller than pre-specified cut-off values. 95% coverage of the population value and its 95% significance.

Note that, concerning (3), the *ppp*-value, which defined as the proportion of chi-square values obtained in the simulated data that exceed that of the actual data and *ppp*-values around 0.50 indicate a well-fitting model.

To determine whether the simulation results resemble a good model fit, the proportion of replications where the critical value of 0.05 is exceeded should be close to 0.05, as *p*-values are expected to be uniformly distributed. The 95% coverage is defined as the percentage of replications for which the 95% CI included the population value of $\Delta M = 0.5$. The significance criterion was defined as the percentage of datasets for which the 95% CI did not include zero, i.e., the percentage of datasets for which we would have concluded that ΔM is larger than zero in the population, which it was for all populations.

RESULTS

Table 4 provides the results for Model 1 and 2 with ML and Bayesian estimation, **Table 5** provides the results for Models 3a–3d and **Table 6** provides the results for Models 4a–4d. We will first discuss the results row wise, i.e., per model, followed by a column wise comparison, i.e., per population.

When full MI (Model 1) is applied to populations where there are differences on the intercepts between the groups (Pop. 2–7) there is a bias in the latent factor means, which does not occur when applied to a population with no differences (Pop. 1). The only exception is Population 5 with many small intercept differences. However, the coverage is smaller than 95% in this case. When partial MI (Model 2) is applied to populations with intercept differences between all intercepts (Pop. 5–7) there is a large bias, which does not occur when applied to populations with only 2 intercepts having differences between the groups (Pop. 2–4) or without any intercept differences (Pop. 1). Applying approximate MI to all intercepts (Model 3) leads to no bias when applied to a population with no differences (Pop. 1), or a population with small differences (Pop. 5). It does lead to a bias in the other populations with moderate or large intercept differences no matter which prior variance was used (Pop. 2,4,6,7). Applying approximate MI to only those intercepts that are different in the population (Model 4 applied to Pop. 2 and 3) leads to a bias, where the magnitude of the bias is dependent on the prior variance specified.

In population 1, with no intercept differences, the bias is smallest for the Model with strict MI, but the coverage is higher for the models with approximate MI and a high precision of the prior (Models 3c and 3d). In the population with 2 small differences, approximate MI with a high precision of the prior (Models 3c and 3d) modestly outperforms strict and partial MI in terms of bias and coverage. For the populations with moderate and large differences, and invariance on either 2 or 4 items, partial MI is clearly the best model. Also, for the model with many small differences, approximate MI with a high precision of the prior (Models 3c and 3d) just outperforms strict MI and clearly outperforms partial MI. The models with a low precision of the prior were never unbiased.

As pointed out by one of the reviewers, comparing **Table 3** and **Table 4** on Population 5, partial MI using both ML and Bayes gave smaller relative bias, smaller standard errors, and more accurate 95% coverage than model 3c and model 3d. Indeed, the coverage of model 3c and 3d is too high because in an ideal situation the

Table 4 | Simulation results for Model 1 and 2.

Model	Outcome	Population 1: No differences		Population 2: 2 items with small differences		Population 3: 2 items with moderate differences		Population 4: 2 items with large differences		Population 5: 4 items with small differences		Population 6: 4 items with moderate differences		Population 7: 4 items with large differences	
		ML	Bayes	ML	Bayes	ML	Bayes	ML	Bayes	ML	Bayes	ML	Bayes	ML	Bayes
#1 Full measurement invariance	Estimated ΔM and *SE*	0.4995 (0.0917)	0.4976 (0.0980)	0.5117 (0.0920)	0.5097 (0.0985)	0.6501 (0.0976)	0.6488 (0.1061)	2.3699 (0.2192)	2.3366 (0.2224)	0.5362 (0.0923)	0.5341 (0.0989)	0.8786 (0.0995)	0.8768 (0.1086)	2.6222 (0.1679)	2.6619 (0.1837)
	Convergence	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Relative bias ΔM%	−0.1	−0.48	2.34	1.94	30.02	29.76	373.98	367.32	7.24	6.82	75.72	75.72	424.44	432.38
	95% coverage	95.9%	95.1%	96.1%	94%	65.3%	56.9%	0%	0%	93.9%	91.2%	2.2%	1.5%	0%	0%
	95% significance	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
#2 Partial measurement invariance	Estimated ΔM and *SE*	0.4990 (0.0979)	0.4863 (0.0988)	0.4990 (0.0979)	0.4863 (0.0988)	0.4990 (0.0979)	0.4863 (0.0988)	0.4990 (0.0979)	0.4863 (0.0988)	0.5298 (0.0984)	0.5169 (0.0992)	0.7904 (0.103)	0.8079 (0.104)	2.0493 (0.148)	1.9829 (0.123)
	Convergence	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
	Relative bias ΔM%	−0.2	−2.74	−0.2	−2.74	−0.2	−2.74	−0.2	−2.74	5.96	3.38	58.08	61.58	309.86	296.58
	95% coverage	94.8%	94.3%	94.8%	94.3%	94.8%	94.3%	94.8%	94.3%	94.6%	94.4%	17.3%	13.9%	0%	0%
	95% significance	99.9%	99.9%	99.9%	99.9%	99.9%	99.9%	99.9%	99.9%	100%	100%	100%	100%	100%	100%

Table 5 | Simulation results for Model 3.

Model	Outcome	Population 1: No differences	Population 2: 2 items with small differences	Population 3: 2 items with moderate differences	Population 4: 2 items with large differences	Population 5: 4 items with small differences	Population 6: 4 items with moderate differences	Population 7: 4 items with large differences
#3a N∼(0, 0.5)	Estimated ΔM and *SE*	0.0404 (0.5161)	0.8537 (0.5923)	0.6417 (0.6627)	1.1153 (0.7033)	0.8779 (0.5924)	0.9018 (0.6975)	2.3347 (0.7101)
	Convergence	100%	100%	100%	100%	100%	99.4%	99.8%
	Relative bias ΔM(%)	−91.92	70.74	28.34	123.06	75.58	80.36	366.94
	95% coverage	92.9%	100%	100%	100%	100%	100%	0%
	95% significance	0%	0.1%	0%	2.1%	0.1%	0%	100%
#3b N∼(0, 0.05)	Estimated ΔM and *SE*	0.4143 (0.2294)	0.5378 (0.2239)	0.6125 (0.2393)	1.1672 (0.2612)	0.5622 (0.2240)	0.8560 (0.2409)	2.3644 (0.2709)
	Convergence	100%	100%	100%	100%	100%	100%	100%
	Relative bias ΔM(%)	−17.14	7.56	22.5	133.44	12.44	71.2	372.88
	95% coverage	100%	100%	99.9%	7.3%	99.9%	87.6%	0%
	95% significance	45.9%	89.5%	97.1%	100%	94.2%	100%	100%
#3c N∼(0, 0.01)	Estimated ΔM and *SE*	0.4554 (0.1246)	0.5124 (0.1352)	0.6167 (0.1320)	1.6506 (0.2169)	0.5368 (0.1353)	0.8596 (0.1368)	2.4984 (0.2011)
	Convergence	100%	100%	100%	100%	100%	100%	100%
	Relative bias ΔM(%)	−8.92	2.48	23.34	230.12	7.36	71.92	399.68
	95% coverage	98.2%	99.7%	94.7%	0%	99.7%	17.2%	0%
	95% significance	99.4%	99.8%	100%	100%	99.9%	100%	100%
#3d N∼(0, 0.005)	Estimated ΔM and *SE*	0.4671 (0.1072)	0.5084 (0.1173)	0.6218 (0.1122)	1.9494 (0.2205)	0.5328 (0.1175)	0.8611 (0.1142)	2.5453 (0.1900)
	Convergence	100%	100%	100%	100%	100%	100%	100%
	Relative bias ΔM(%)	−6.58	1.68	24.36	289.88	6.56	72.22	409.06
	95% coverage	97.3%	98.9%	86.9%	0%	98.6%	7.7%	0%
	95% significance	99.8%	100%	100%	100%	100%	100%	100%

95% confidence interval should cover the true parameter value in exactly 95% of the times. The coverage of almost 100% is probably caused by the standard error in model 3c to be overestimated, which can result in the reduction of statistical power. In conclusion, approximate MI should not be applied when full MI holds in the population. If large differences exist in the population on only a few intercepts, partial MI outperforms approximate MI, but partial approximate MI with a large prior variance can also be used. If moderate or small differences exist in the population on only a few intercepts, partial approximate MI is preferred. If small differences exist in the population on many intercepts, approximate MI outperforms applying full MI.

RESOLVING THE ALIGNMENT ISSUE

In the previous section we have seen that some of the parameter values, in our case the difference between the latent means, that generated the data are not recovered due to the alignment problem, which reflects an indeterminacy in the CFA. Applying approximate invariance using the DIFF statement tends to pull the deviating parameter toward the average of the parameters across all groups. As a result the deviating parameter will be underestimated and the invariant parameters overestimated, see also the simulation results in Asparouhov and Muthén (2013). With biased intercepts the latent factor means will be biased as well and this is what we call the alignment issue (Asparouhov and Muthén, 2013; in preparation). If one would use plain BSEM the results of the CFA model might be biased in the estimates of the latent mean difference scores, especially when the precision of the DIFF prior is low (i.e., large prior variance), which is undesirable. There are two options to deal with the alignment issues: (1) Freeing the parameters found not invariant (as in Asparouhov and Muthén, 2013), or (2) using the alignment methods available in Mplus v7.1. In the current paper we will focus on the second option, for a comparison of both methods see Asparouhov and Muthén (in preparation for the special issue).

In Mplus v7.1 the alignment-method handles the issue of alignment through rotation. The rotation for the alignment-method uses the same principles as for EFA (Jennrich, 2006) and

Table 6 | Simulation results for Model 4.

Model	Outcome	Population 2: 2 items with small differences	Population 3: 2 items with moderate differences	Population 4: 2 items with large differences
#4a (N∼(0, 0.5))	Estimated ΔM and *SE*	0.4926 (0.0993)	0.4939 (0.0994)	0.4998 (0.1000)
	Convergence	100%	100%	100%
	Relative bias ΔM(%)	−1.48	−1.22	−0.04
	95% coverage	95%	95%	95.5%
	95% significance	99.9%	99.9%	99.9%
#4b (N∼(0, 0.05))	Estimated ΔM and *SE*	0.4931 (0.0985)	0.5051 (0.0996)	0.5703 (0.1072)
	Convergence	100%	100%	100%
	Relative bias ΔM(%)	−1.38	1.02	14.09
	95% coverage	95.7%	95.6%	90.9%
	95% significance	99.9%	99.9%	100%
#4c (N∼(0, 0.01))	Estimated ΔM and *SE*	0.4952 (0.0966)	0.5403 (0.0999)	1.4388 (0.2410)
	Convergence	100%	100%	100%
	Relative bias ΔM(%)	−0.96	8.06	187.76
	95% coverage	96.4%	93.6%	3.6%
	95% significance	100%	100%	100%
#4d(N∼(0, 0.005))	Estimated ΔM and *SE*	0.4971 (0.0954)	0.5656 (0.0999)	1.9635 (0.2390)
	Convergence	100%	100%	100%
	Relative bias ΔM(%)	−0.58	13.12	292.7
	95% coverage	96.4%	91.4%	0%
	95% significance	100%	100%	100%

is described in more details in Asparouhov and Muthén (2013). As stated in the version 7.1 Mplus language addendum (Muthén and Muthén, 2013, p. 2): "the alignment optimization method consists of three steps:

(1) Analysis of a configural model with the same number of factors and same pattern of zero factor loadings in all groups.
(2) Alignment optimization of the measurement parameters, factor loadings and intercepts/thresholds according to a simplicity criterion that favors few non-invariant measurement parameters.
(3) Adjustment of the factor means and variances in line with the optimal alignment."

The third step in this procedure is expected to decrease the bias in the latent variable means as we discussed above. We included the syntax `ANALYSIS: ALIGNMENT = FIXED (BSEM);` where `FIXED` enforces the first latent mean to be zero and the second latent mean to be estimated. When `FREE` would have been specified all latent means would have been estimated, which is only recommended if more than two groups are specified (Asparouhov and Muthén, 2013. p. 16). Furthermore, `BSEM` refers to the combination of the alignment-method with the `DIFF` statements.

To explore the performance of the BSEM-alignment method we ran additional models on population 5 where groups exhibit small differences on the intercepts of all four items (see **Table 3**). Recall that the bias for this population when applying plain BSEM was 7.36% ($SE = 0.1353$) with the DIFF statement imposed upon all intercepts, but where the factor loadings were constrained across groups (denoted by Model 5a). When population 5 was confronted with a model that imposed plain-BSEM with the DIFF statement on both intercepts and factor loadings (Model 5b) we encountered a bias of 3.62% ($SE = 0.1279$). When the ALIGNMENT = FIXED(BSEM) command was added on top of DIFF statements (Model 5c) the bias appeared to be 4.28% with a lower *SE* of 0.1174. Thus, in this situation the alignment method leads to less bias. Note that these findings are all conditional on normal priors for the DIFF statements with a prior variance of 0.01.

Since prior variance turned out to influence bias and SE's in previous runs we ran Model 5b and Model 5c with prior variances of 0.5 and 0.05 in the DIFF statements. Model 5b with a prior variance of 0.05 yielded a bias of −1.35% ($SE = 0.2039$) and Model 5c yielded a bias of 4.02% ($SE = 0.1413$). Just as with a prior variance of 0.01, if the ALIGNMENT command is added to the DIFF commands, the standard error decreased. When the prior variance of the DIFF statements is increased to 0.5 Model 5b yielded a bias of −43.46% ($SE = 0.4035$), whereas Model 5c with the alignment method performed much better in terms of relative bias and *SE* : −1.34% and a *SE* of 0.1413.

Similar findings were obtained for a population with large differences on all intercepts across groups (Population 7). For this population the bias and SE's were even higher: 399%

($SE = 0.2011$), 391% ($SE = 0.1940$) and 394% ($SE = 0.3057$) for Models 5a, 5b, and 5c with prior variances of 0.01, respectively. It appeared the alignment method, just like plain BSEM, does not resolve the incurred bias when group intercept differences are moderate or high, especially if many items are affected. Since we only used four items in our simulation design more research is needed to investigate whether it is the magnitude of the non-invariance or the number of items affected.

Finally, we ran simulations with models 5a, 5b and 5c for the populations where only two of the items in the population were dissimilar (Populations 1–3, see **Table 3**), again with a prior variance of 0.01 for the DIFF priors. The results were comparable with the results discussed above (results not shown but these can be requested from the first author). With the ALIGNMENT command we obtained slightly smaller SEs with only small differences in the population compared to approximate MI without the ALIGNMENT command. However, with moderate or large intercept differences between groups the bias and SE for all models were once more too high.

Taken together, DIFF statements imposed upon parameters without the support of an ALIGNMENT command (i.e., plain BSEM) introduced slightly higher standard errors compared to DIFF statements that are combined with the ALIGNMENT command.

CONCLUSION

If a researcher wants to compare latent means across groups or over time one has four options:

(1) Impose (full or) scalar MI. When a full MI structure results in approporiate model fit any difference in latent means respresents true, unbiased difference between groups/timepoints.
(2) Impose partial MI, where one studies the size of the differences between unconstrained loadings and/or intercepts, and constrains all loadings and intercepts except for the one loading/intercept with the largest difference, which is released. If the fit statistics are satisfied, any difference in the latent means is indicative of true mean differences. Sumscores, however, are biased due to the items where differences in the intercepts/factor loadings are allowed (Steinmetz, 2013).
(3) Impose no MI, leading to the conclusion that the latent means cannot be used for comparing groups because any difference in the latent means can be caused by many differences.

With Muthén and Asparouhov's introduction of approximate MI (2012a; 2012b; 2013), a fourth option for testing MI became available.

(4a) Approximate MI salvages MI in the case of seemingly ignorable (i.e., near zero) differences between parameters.

Or when combined with partial MI:

(4b) Partial approximate MI, which is a hybrid form of partial MI and approximate MI.

The results of our paper have shown that applying approximate MI might provide a safe passage through the narrow channel between Italy and Sicily in order to facilitate the escape from the mythical sea monsters Scylla and Charybdis, just as Odysseus was able to. The whirlpools caused by Charybdis, who dislikes comparing latent mean scores if the factor loadings/intercepts are dissimilar across groups, can be avoided. The reason is that with approximate MI, parameters are restricted to be closer to each other than with partial MI. The use of Bayesian statistics on the difference in parameters introduces a posterior distribution, which tries to find a compromise between the ideal situation (difference = zero) and the situation we find in the data. The willingness to compromise between ideology and reality has the following effect: the posterior difference in parameters across groups is close enough to its ideal zero to allow latent mean comparisons, yet close enough to the reality of the data to result in acceptable model fit. As was noted by one of the reviewers, a crucial distinction between partial invariance factor models and the Bayesian approach involving priors is that the former typically is coupled with a substantive interpretation of the group differences in the parameters of interest. Although substantive considerations may certainly help inform the nature of group differences, there is always a risk of *ad hoc* reasoning in applications. The Bayesian approach may do more justice to the unexpected and possibly inexplicable failures of invariance. In a related vein, partial measurement models have often been criticized for lacking specificity in the sense that large modification indices of certain items/indicators may actually reflect failures of invariance of other items/indicators (see e.g., Reise et al., 1993).

Likewise it is possible to avoid Scylla, who will devour badly fitting models resulting from forcing scalar MI on a model where differences do exist. Both our empirical example and the simulation study have taught us that there seems to be an optimum specification of the prior variance. The alignment method provides promising results for decreasing the influence of the prior specifications, but more research is warranted.

We recommend the following procedure if the test for full MI fails. First, determine which parameters are different between groups, for example by using modification indices or by using the DIFFERENCE OUTPUT which is obtained when the DIFF statement is used in Mplus. The latter output provides each parameter with a significance test for its deviance with its constrained counterpart. Do not impose MI when there are large parameter differences across groups, or impose approximate MI when you are able to locate just a few deviating parameters. If there are (many) small differences we recommend to apply full approximate MI. Use the ALIGNMENT method when you don't want to use small prior variances in the DIFF statements. We acknowledge the issue of defining "small differences." With "small" we imply that parameters of substantive interest do not change in a meaningful way if MI does not fully hold (cf. Oberski, 2013). We note that the choice of the priors is extremely important and since the field of approximate MI is rather unexplored we advise always do sensitivity analyses and never just "choose" a prior value. One aspect influencing the definition of a "small difference" is that the prior are sensible for a given data set, and hence, that the choice of the prior variance does have huge

implications on the parameter estimates. In particular, because the difference in intercepts is a function of the scaling of the observed variable, as was noted by one of the reviewers, it may be helpful to relate the variances of the normal priors to the scaling (or variability) of the observed variables. For example, for a prior with hyperparameters N (0, 0.01) indicates there is a (subjective belief of) 95% chance that the absolute intercept difference is equal or smaller than 0.01 [i.e., sqrt(0.01) = 0.1].

Since the field of approximate MI is relatively new we propose the following research agenda. First, there are two variables influencing the performance of MI: (1) the number of items with differences on the factor loading or intercepts and (2) the size of the difference itself. What we do not know is what the exact cut-off values are for these decisions (number of items and magnitude of differences). This topic needs further attention, given that it can help researchers make informed choices about applying partial or approximate MI without having to test them both. Second, more simulation studies have to be performed to find out which prior specification in which model is to be advised, since the optimum prior specification is model and data dependent. Third, the bias in substantive results if the incorrect type of MI is used should be investigated in more detail. Fourth, more research is needed to study the effects of the alignment method. Fifth, misspecification of the baseline model should be further investigated. A fifth area for further exploration is the comparison of the approximate MI procedure with alternative approaches, for example the commonly used delta-goodness-of-fit-indexes (i.e., ΔGFI; Cheung and Rensvold, 2002; Chen, 2007). And finally, in our simulation study we used a relative large sample size in relation to the degrees of freedom. It should be investigated which sample sizes vis-à-vis model DFs are needed for the Bayesian analysis to work properly. It is expected that the Bayesian test for MI can deal with smaller sample sizes compared to the ML counterparts, as was also the case for regular SEM models, see Lee and Song (2004) and Van de Schoot et al. (submitted).

It should be noted that approximate MI might be an interesting alternative approach for testing MI, but it does not replace the original MI test which is based on, for example chi-square difference testing. Approximate MI, as introduced in our paper provides a first step in this challenging and promising new area of testing and exploring MI if the chi-square test, or any other test, rejects the invariance model. Also, our paper provides a warning not to use approximate MI in all situations where MI is tested, but this warning message also applies to strict MI and partial MI.

ACKNOWLEDGMENTS

The first author was supported by a grant from the Netherlands organization for scientific research: NWO-VENI-451-11-008.

REFERENCES

Asparouhov, T., and Muthén, B. (2013). *Multiple Group Factor Analysis Alignment.* Available online at: http://statmodel.com/examples/webnotes/webnote18.pdf. Mplus Web Notes: No. 18. [May 30, 2013].

Ajzen, I. (1991). The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* 50, 179–211. doi: 10.1016/0749-5978(91)90020-T

Busse, R., Geissler, A., and Quentin, W., Wiley, M., (eds.). (2011). *Diagnosis Related Groups in Europe. Moving Towards Transparency, Efficiency and Quality in Hospitals*. New York, NY: Open University Press.

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling* 14, 464–504. doi: 10.1080/10705510701301834

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equ. Modeling* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Cohen, J. (1992). A power primer. *Psychol. Bull.* 112, 155–159. doi: 10.1037/0033-2909.112.1.155

De Boeck, P. (2008). Random item IRT models. *Psychometrika* 73, 533–559. doi: 10.1007/s11336-008-9092-x

Hoogland, J. J., and Boomsma, A. (1998). Robustness studies in covariance structure modeling: an overview and a meta-analysis. *Socio. Meth. Res.* 26, 329–367. doi: 10.1177/0049124198026003003

Jennrich, R. I. (2006). Rotation to simple loadings using component loss function: the orthogonal case. *Psychometrika*, 71, 173–191. doi: 10.1007/s11336-003-1136-B

Metselaar, E. E. (1997). *Assessing the Willingness to Change: Construction and Validation of the DINAMO*. Doctoral dissertation, Free University of Amsterdam.

Kruschke, J. K., Aguinis, H., and Joo, H. (2012). The time has come! Bayesian methods for data analysis in the organizational sciences. *Organ. Res. Methods* 15, 722–752. doi: 10.1177/1094428112457829

Lee, S.-Y., and Song, X.-Y. (2004). Evaluation of the bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivar. Behav. Res.* 39, 653–686. doi: 10.1207/s15327906mbr3904_4

Mellenbergh, G. J. (1989). Item bias and item response theory. *Int. J. Educ. Res.* 13, 127–143. doi: 10.1016/0883-0355(89)90002-5

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

Morren, M., Gelissen, J. P. T. M., and Vermunt, J. (2011). Dealing with extreme response style in cross-cultural research: a restricted latent class factor analysis approach. *Sociol. Methodol.* 41, 13–47. doi: 10.1111/j.1467-9531.2011.01238.x

Muthén, B., and Asparouhov, T. (2012a). *New Features in Mplus v7 Lecture 3*. Available online at: http://mplus.fss.uu.nl/2012/09/12/.the-workshop-new-features-of-mplus-v7/ [Accessed November 6, 2012].

Muthén, B., and Asparouhov, T. (2012b). Bayesian SEM: a more flexible representation of substantive theory. *Psychol. Methods* 17, 313–335. doi: 10.1037/a0026802

Muthén, B., and Asparouhov, T. (2013). *BSEM Measurement Invariance Analysis. Mplus Web Notes: No. 17*. Available online at: www.statmodel.com

Muthén, L. K., and Muthén, B. O. (1998-2012). *Mplus User's Guide. 7th* Edn. Los Angeles, CA: Muthén and Muthén.

Muthén, L. K., and Muthén, B. O. (2013). *Version 7.1 Mplus Language Addendum.* Available online at: www.statmodel.com 31-05-2013.

Neukrug, E. (2011). *The World of the Counselor: An Introduction to the Counseling Profession*. Belmont, CA: Thomson Brooks/Cole.

Oberski, D. L. (2013). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*. doi: 10.1093/pan/mpt014

Palm, I., Leffers, F., Emons, T., Van Egmond, V., and Zeegers, S. (2008). *De GGz ontwricht: Een praktijkonderzoek naar de gevolgen van het nieuwe zorgstelsel in de geestelijke gezondheidszorg (Problems in mental health care: An applied research on the impact of the Health Insurance Law in mental health care)*. Den Haag: SP.

Rensvold, R. B., and Cheung, G. W. (2001). "Testing for metric invariance using structural equation models: Solving the standardization problem," in *Research in Management:* Vol. 1 *Equivalence in Measurement* eds C. A. Schriesheim and L. L. Neider (Greenwich, CT: Information Age), 21–50.

Reise, S. P., Widaman, K. F., and Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychol. Bull.* 114, 552–566. doi: 10.1037/0033-2909.114.3.552

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

Smullen, A. (2013). "Institutionalizing professional conflicts through financial reforms: The case of DBC's in dutch mental healthcare," in *Professionals Under Pressure: The Reconfiguration of Professional Work in Changing Public Services* eds M. Noordegraaf and A. J. Steijn (Amsterdam: Amsterdam University Press), 92–109.

Steenkamp, J. M., and Baumgartner, H. (1998). Assessing measurement invariance in cross national consumer research. *J. Consum. Res.* 25, 78–90. doi: 10.1086/209528

Steinmetz, H. (2013). Analyzing observed composite differences across groups. Is partial measurement invariance enough? *Methodology* 9, 1–12. doi: 10.1027/1614-2241/a000049

Tummers, L. G., Steijn, A. J., and Bekkers, V. J. J. M. (2012). Explaining willingness of public professionals to implement public policies: content, context, and personality characteristics. *Public Adm.* 90, 3, 716–736. doi: 10.1111/j.1467-9299.2011.02016.x

Tummers, L. G. (2012). Policy alienation public professionals: the construct and its measurement. *Public Adm. Rev.* 72, 516–525. doi: 10.1111/j.1540-6210.2011.02550.x

van Zuiden, M., Heijnen, C. J., van de Schoot, R., Amarouchi, K., Maas, M., Vermetten, E., et al. (2011). Cytokine production by leukocytes of military personnel with depressive symptoms after deployment to a combat-zone: a prospective, longitudinal study, *PLoS ONE* 6:e29142. doi: 10.1371/journal.pone.0029142

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., and van Aken, M. A. G. (2013). A gentle introduction to bayesian analysis: applications to developmental research. *Child Dev.* doi: 10.1111/cdev.12169

Van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *Eur. J. Dev. Psychol.* 9, 486–492. doi: 10.1080/17405629.2012.686740

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organ. Res. Methods* 5, 139–158. doi: 10.1177/1094428102005002001

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 18 June 2013; accepted: 30 September 2013; published online: 23 October 2013.

*Citation: van de Schoot R, Kluytmans A, Tummers L, Lugtig P, Hox J and Muthén B (2013) Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. Front. Psychol. **4**:770. doi: 10.3389/fpsyg.2013.00770*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

APPENDIX A

DATA:
FILE IS data Tummers.dat;

VARIABLE:
NAMES ARE CaseNR BCPsych Veran1 Veran2 Veran3 Veran4 mean;
USEVARIABLES ARE Veran1 Veran2 Veran3 Veran4;
MISSING ARE ALL (-9999);
KNOWNCLASS IS g (BCPsych=0 BCPsych=1);
CLASSES IS g(2);

ANALYSIS:
MODEL IS allfree;
TYPE IS is mixture;
ESTIMATOR IS BAYES;
Bconvergence=0.01;
Biterations = 500000 (20000);
processor is 8;
chains is 8;
bseed 100;

MODEL:
%overall%
Willingness by Veran1-Veran4*(1-4);
Willingness@1;
[Willingness@0];
[Veran1- Veran4] (nu#_1 - nu#_4);

MODEL PRIORS:
DO(1, 4) DIFF(nu1_#-nu2_#) ∼ N(0, 0.5);

OUTPUT:
SAMPSTAT TECH1 TECH8 STAND(STDYX);

PLOT:
type is plot2;

APPENDIX B

The population input file for the population 3:

MODEL POPULATION:
f1 by y1@.7 y2@.6 y3@.4 y4@.2;
f1@1; [f1@0];
y1-y4@1;
[y1@0]; [y2@0];
[y3@-.1]; [y4@-.1];

MODEL POPULATION-g2:
f1 by y1@.7 y2@.6 y3@.4 y4@.2;
f1@1; [f1@.5];
y1-y4@1;
[y1@0]; [y2@0]; [y3@.1]; [y4@.1]

frontiers in PSYCHOLOGY

METHODS ARTICLE
published: 23 September 2014
doi: 10.3389/fpsyg.2014.01022

Measurement bias detection with Kronecker product restricted models for multivariate longitudinal data: an illustration with health-related quality of life data from thirteen measurement occasions

***Mathilde G. E. Verdam*[1,2]* and *Frans J. Oort*[1,2]**

[1] Department of Medical Psychology, Academic Medical Centre, University of Amsterdam, Amsterdam, Netherlands
[2] Faculty of Social and Behavioral Sciences, Research Institute Child Development and Education, University of Amsterdam, Amsterdam, Netherlands

Edited by:
Peter Schmidt, Higher School of Economics International Scientific-Educational Laboratory of Socio-Cultural Research, Russia

Reviewed by:
Christian Geiser, Utah State University, USA
A. Klein, Goethe University, Germany

***Correspondence:**
Mathilde G. E. Verdam, Department of Child Development and Education, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, Netherlands
e-mail: m.g.e.verdam@uva.nl

Highlights

- Application of Kronecker product to construct parsimonious structural equation models for multivariate longitudinal data.
- A method for the investigation of measurement bias with Kronecker product restricted models.
- Application of these methods to health-related quality of life data from bone metastasis patients, collected at 13 consecutive measurement occasions.
- The use of curves to facilitate substantive interpretation of apparent measurement bias.
- Assessment of change in common factor means, after accounting for apparent measurement bias.

Longitudinal measurement invariance is usually investigated with a longitudinal factor model (LFM). However, with multiple measurement occasions, the number of parameters to be estimated increases with a multiple of the number of measurement occasions. To guard against too low ratios of numbers of subjects and numbers of parameters, we can use Kronecker product restrictions to model the multivariate longitudinal structure of the data. These restrictions can be imposed on all parameter matrices, including measurement invariance restrictions on factor loadings and intercepts. The resulting models are parsimonious and have attractive interpretation, but require different methods for the investigation of measurement bias. Specifically, additional parameter matrices are introduced to accommodate possible violations of measurement invariance. These additional matrices consist of measurement bias parameters that are either fixed at zero or free to be estimated. In cases of measurement bias, it is also possible to model the bias over time, e.g., with linear or non-linear curves. Measurement bias detection with Kronecker product restricted models will be illustrated with multivariate longitudinal data from 682 bone metastasis patients whose health-related quality of life (HRQL) was measured at 13 consecutive weeks.

Keywords: Kronecker product, multivariate longitudinal data, measurement invariance, structural equation modeling (SEM), longitudinal three-mode model (L3MM), health-related quality of life (HRQoL)

A valid assessment of change requires that the meaning of the construct stays the same across measurement occasions (Meredith, 1993). Longitudinal measurement invariance is usually investigated with the longitudinal factor model (LFM). When R latent variables are measured with K observed variables at J measurement occasion, the mean, and covariance structures are given by:

$$\boldsymbol{\mu} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa}, \tag{1}$$

and:

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{2}$$

where $\boldsymbol{\tau}$ is a JK-vector of intercepts, $\boldsymbol{\Lambda}$ is a $JK \times JR$ matrix of common factor loadings, $\boldsymbol{\kappa}$ is a JR-vector of common factor means, $\boldsymbol{\Phi}$ is a $JR \times JR$ symmetric matrix containing the variances and covariances of the common factors, and $\boldsymbol{\Theta}$ is a $JK \times JK$ symmetric matrix containing the variances and covariances of the residual factors. Although the LFM can be used to model multiple measurement occasions, these models become progressively large and unmanageable when the number of occasions increases.

One solution to this problem is the imposition of Kronecker product restrictions that profit from the three-mode structure of multivariate longitudinal data (Oort, 2001). The modes refer to

the variables, the measurement occasions and the subjects, and the resulting longitudinal three-mode models (L3MMs) are more parsimonious and have attractive interpretation. For example, Kronecker product restrictions can be imposed on factor loadings and intercepts to comply with measurement invariance, using:

$$\mathbf{\Lambda} = \mathbf{I} \otimes \mathbf{\Lambda}_0, \quad (3)$$

and:

$$\boldsymbol{\tau} = \mathbf{u} \otimes \boldsymbol{\tau}_0, \quad (4)$$

where $\mathbf{\Lambda}_0$ is a $K \times R$ matrix of invariant common factor loadings, $\boldsymbol{\tau}_0$ is a $K \times 1$ vector of invariant intercepts, $\mathbf{I}$ is a $J \times J$ identity matrix, $\mathbf{u}$ is a $J \times 1$ vector of ones, and the symbol $\otimes$ denotes the Kronecker product. These restrictions imply that factor loadings $\mathbf{\Lambda}_0$ and intercepts $\boldsymbol{\tau}_0$ apply to all measurement occasions. Although Kronecker product restrictions are convenient to model measurement invariance, they require special methods for the investigation of violations of measurement invariance (i.e., measurement bias).

Specifically, to detect measurement bias in Kronecker product restricted models, we introduce additional matrices $\mathbf{A}$ and $\mathbf{B}$ to accommodate possible violations of measurement invariance, using:

$$\mathbf{\Lambda} = \mathbf{I} \otimes \mathbf{\Lambda}_0 + \mathbf{A}, \quad (5)$$

and:

$$\boldsymbol{\tau} = \mathbf{u} \otimes \boldsymbol{\tau}_0 + \mathbf{B}. \quad (6)$$

These additional matrices consist of measurement bias parameters that are either fixed at zero or free to be estimated. This method thus enables the detection of measurement bias in individual parameters of $\mathbf{\Lambda}$ and $\boldsymbol{\tau}$. In this way, it is possible to establish partial measurement invariance (Byrne et al., 1989). Moreover, in cases of measurement bias, it is also possible to model the bias over time, e.g., with linear or non-linear curves, which can facilitate interpretation.

The aim of the present paper is to illustrate the detection of measurement bias with Kronecker product restricted models using multivariate longitudinal data from 682 bone metastasis patients whose health-related quality of life (HRQL) was measured in 13 consecutive weeks.

METHODS

Patients with painful bone metastases from a solid tumor were enrolled from 17 radiotherapy institutes in The Netherlands. Patients were randomized to receive radiotherapy of a single fraction vs. multiple fractions as palliative treatment for pain. Inclusion criteria were having one or more painful bone metastases treatable in one target volume and having a pain score of at least 2 on an 11-point scale from 0 (no pain at all) to 10 (worst imaginable pain) at time of admission to the radiotherapy. Exclusion criteria were having metastases of malignant melanoma or renal cell carcinoma, having metastases in the cervical spine, having previously been irradiated for the same metastases, or having a pathological fracture that needed surgical fixation or a spinal cord compression. Side effects from radiation therapy vary depending on the part of the body being treated, and may include skin changes (dryness, itching, peeling, or blistering), fatigue, loss of appetite, hair loss, diarrhea, nausea, and vomiting. Most of these side effects go away within a few weeks after radiation therapy.

HRQL questionnaires were administered at 13 measurement occasions, before (T0) and every week after treatment with radiotherapy (T1 through T12). These data are a subset of data from the Dutch Bone Metastasis Study (Steenland et al., 1999; Van der Linden et al., 2004). For the current study only patients who survived at least 13 weeks from the start of treatment were included, which resulted in a total sample size of 682 patients (354 women). Patients' primary tumor was either breast cancer ($n = 321$), prostate cancer ($n = 181$), lung cancer ($n = 106$), or other ($n = 74$). Ages ranged from 33 to 90, with a mean of 64.2 (standard deviation 11.5).

Treatment progression, therapeutic effects and/or side effects may influence patients' health status. In the area of HRQL a theoretical framework of measurement bias has been developed which describes how changes in patients' health status may prompt behavioral, cognitive, and affective processes that affect patients' response tendencies (Sprangers and Schwartz, 1999). Therefore, it seems worthwhile to investigate measurement bias in our sample of bone metastases patients.

MEASURES

HRQL was assessed with multiple questionnaires (for more information see Verdam et al., submitted). Forty-five Items were grouped using confirmatory factor analyses and substantive considerations to compute eight health-indicators: physical functioning (PF; 4 items), mobility (MB; 5 items), social functioning (SF; 2 items), depression (DP; 8 items), listlessness (LS; 6 items), pain (PA; 4 items), sickness (SI; 6 items), and treatment related symptoms (SY; 11 items). All scale scores were calculated as mean item scores, ranging from 1 to 4, with higher scores indicating more symptoms or dysfunctioning.

Intermittent missing item- and scale scores were imputed using expectation-maximization. Per assessment, 29–35% of respondents showed missing item scores and 1–3% of respondents showed intermittent missing scale scores. Cronbach's alpha coefficients indicated moderate to good internal consistency reliability (PF, alpha = 0.93; MB, alpha = 0.91; SF, alpha = 0.80; DP, alpha = 0.94; LS, alpha = 0.72; SI, alpha = 0.74; PA alpha = 0.74; SY, alpha = 0.69).

STRUCTURAL EQUATION MODELING

Structural equation models were fitted to the means, variances and covariances of the eight observed health indicators using OpenMx (Boker et al., 2011). OpenMx syntax is available in Appendix I[1]. To achieve identification of all model parameters,

[1]OpenMx was used for statistical analyses because it provides a matrix algebraic approach to structural equation modeling that facilitates the decomposition of matrices that is required for the imposition of Kronecker product restrictions. Other statistical software (e.g., LISREL and Mplus) could also be used for statistical analyses presented in this article, but these programs require a much longer, more complicated script as they only allow inequality constraints on individual parameters.

scales and origins of the common factors were established by fixing the factor means at zero and the factor variances at one. When factor loadings and intercepts were constrained to be equal across occasion, only first occasion factor means and variances were fixed. Model parameters of the additional matrices **A** and **B** can be freely estimated, with the restriction that the computed matrices of factor loadings and intercepts do not violate the general guidelines for identification as described above. Identification of model parameters of matrices that feature in the Kronecker product restriction imposed on residual factor variances and covariances was achieved by using the guidelines described by Oort (2001).

Detection of measurement bias

The structural equation modeling procedure for the detection of measurement bias included the following steps: (1) establishing an appropriate measurement model, (2) fitting a model of measurement invariance, (3) detection of measurement bias, (4) modeling the bias that was detected, and (5) assessment of change.

Step 1: measurement model. The *Measurement Model* was established on the basis of results of exploratory factor analyses and substantive considerations. To take into account the multivariate longitudinal structure of the data, the longitudinal three-mode model (L3MM; Oort, 2001) was applied. To reduce the complexity of the model (i.e., the number of parameter estimates) Kronecker product restrictions were imposed on residual variances and covariances, using $\boldsymbol{\Theta} = \boldsymbol{\Theta}_{\mathrm{T}} \otimes \boldsymbol{\Theta}_{\mathrm{V}}$. This restriction entails that the matrix of residual variances and covariances ($\boldsymbol{\Theta}$) is estimated indirectly by using a symmetric matrix that contains the relationships between measurement occasions ($\boldsymbol{\Theta}_{\mathrm{T}}$, of dimensions 13×13; with $\boldsymbol{\Theta}_{\mathrm{T}(1,1)} = 1$ for identification purposes) and a diagonal matrix that contains the residual variances of only one measurement occasion ($\boldsymbol{\Theta}_{\mathrm{V}}$, of dimension 8×8). This implies that the changes in residual factor variances and covariances across occasions are proportionate for all residual factors (for more details see Verdam et al., submitted). The *Measurement Model* has no equality constraints across occasions.

Step 2: measurement invariance model. The assumption of longitudinal measurement invariance entails that factor loadings and intercepts are constrained to be equal across occasions. These restrictions were imposed using the Kronecker product with Equations (3) and (4), yielding the *Measurement Invariance Model*. To test the assumption of measurement invariance the model fit of the more restricted model can be compared to the model fit of the model with no equality constraints across occasions. When there is no significant deterioration in model fit, the assumption of measurement invariance can be retained.

Step 3: partial measurement invariance model. Detection of measurement bias was done using a step-by-step modification of the *Measurement Invariance Model*, to yield the *Partial Measurement Invariance Model* which included all occurrences of measurement bias. Measurement bias was operationalized as differences across measurement occasions in parameter estimates of factor loadings or intercepts. An iterative procedure was used, where each invariant factor loading and intercept was investigated one-by-one. Using Equations (5) and (6) all measurement bias parameters across occasions that were associated with one invariant parameter were freely estimated. The violations of measurement invariance that yielded the largest improvement in model fit were incorporated in the model. To test whether partial measurement invariance is tenable the model fit of this model can be compared to the model fit of the model with no equality constraints across measurement occasions. When there is no significant deterioration in model fit, the assumption of partial measurement invariance can be retained. The final model, the *Partial Measurement Invariance Model*, thus includes measurement invariance restrictions on most, but not all, factor loading and intercept parameters.

Step 4: modeling occurrences of measurement bias. In case of measurement bias, the bias was modeled using linear or non-linear curves. The measurement bias parameters were modeled as a function of the time of measurement (using a time-coding), yielding estimates of intercept- and slope-parameters that describe the trend of the bias. When the model fit of the more restricted model did not significantly deteriorate compared to the model fit of the model with freely estimated measurement bias parameters, we retained the model which describes the bias using a linear or non-linear curve. Interpretation of parameter estimates provides insight in the trend of the bias that was detected.

Step 5: assessment of change. Change in the common factor means was assessed in the model where all measurement biases were taken into account. A test of invariance was used to investigate differences in common factor means across occasions. To evaluate the impact of measurement bias on the assessment of change, we inspected the trajectories of common factor means, before and after taking into account measurement bias.

Evaluation of model fit

To evaluate goodness-of-fit the chi-square test of exact fit (CHISQ) was used, where a significant chi-square indicates a significant difference between model and data. However, in the practice of structural equation modeling, exact fit is rare, and with large sample sizes and large numbers of degrees of freedom the chi square test generally turns out to be significant. Therefore, we also considered alternative measures of fit. The root mean square error of approximation (RMSEA; Steiger and Lind, 1980; Steiger, 1990) was used as a measure of approximate fit, where RMSEA values below 0.05 indicate "close" approximate fit and values below 0.08 indicate "reasonable" approximate fit (Browne and Cudeck, 1992). Additionally, the expected cross-validation index (ECVI; Browne and Cudeck, 1989) was used to compare different models for the same data, where the model with the smallest ECVI indicates the model with the best fit. For both the RMSEA and ECVI, 95% confidence intervals were calculated using the program NIESEM (Dudgeon, 2003).

To evaluate differences between hierarchically related models the chi-square difference test ($CHISQ_{diff}$) was used, where a significant chi-square difference indicates a significant difference in model-fit. The chi square difference test applied to hierarchically nested models has essentially the same strengths and weaknesses as the chi square test applied to a single model. Therefore, we additionally considered the ECVI difference test ($ECVI_{diff}$), where the deterioration in model fit of the more restricted model is significant when the value of the ECVI difference is significantly larger than zero.

RESULTS

MEASUREMENT MODEL

Eight health-indicators were modeled to be reflective of two common factors: functional limitations and health impairments (see **Figure 1**). Functional limitations was measured by three observed variables, health impairments was measured by six observed variables, with one observed variable in common. The squares represent observed variables (scale scores), the circles on the top represent the common factors, and the circles on the bottom represent residual factors.

Classification of the common factors was based on the International Classification of Functioning, Disability and Health (World Health Organization, 2002) that provides a framework for the description of health and health-related states. In this framework, the term functioning refers to all body functions, activities and participation, while disability refers to impairments, activity limitations and participation restrictions. These concepts are reflected in the two common factors functional limitations (e.g., limitations of bodily functioning) and health impairments (e.g., health restrictions or symptoms). As social functioning is also considered to be an important factor of HRQL, this scale was added to the measurement and modeled to be influenced by both functional limitations and health impairments (which agrees with participation being a factor of both functioning and disability in the WHO framework).

The *Measurement Model* yielded a chi-square test of exact fit that was significant but the RMSEA measure indicated close fit (see Model 1, **Table 1**).

DETECTION OF MEASUREMENT BIAS

To test the assumption of longitudinal measurement invariance, factor loadings and intercepts were held invariant across occasions using the Kronecker product restriction. The overall fit of the *Measurement Invariance Model* showed reasonable fit (RMSEA = 0.037, see **Table 1**), but comparison with the fit of the model with no across occasions equality constraints showed a significant deterioration in fit [$CHISQ_{diff}(156) = 735.2$, $p < 0.001$; $ECVI_{diff} = 0.54$, 95% CI: 0.39–0.71]. This indicates a violation of measurement invariance.

To investigate which of the equality constraints across occasions on factor loadings and intercepts were not tenable, an iterative measurement bias detection approach was used. Step by step modification of the *Measurement Invariance Model* yielded the *Partial Measurement Invariance Model*, which showed three cases of measurement bias. Each of the measurement biases that was detected will be explained in more detail below. The fit of the *Partial Measurement Invariance Model* was good (RMSEA = 0.035, see **Table 1**), and significantly better than the fit of the *Measurement Invariance Model* [$CHISQ_{diff}(36) = 511.7$, $p < 0.001$; $ECVI_{diff} = 0.63$, 95% CI: 0.50–0.77]. Moreover, comparison with the *Measurement Model* showed that although there

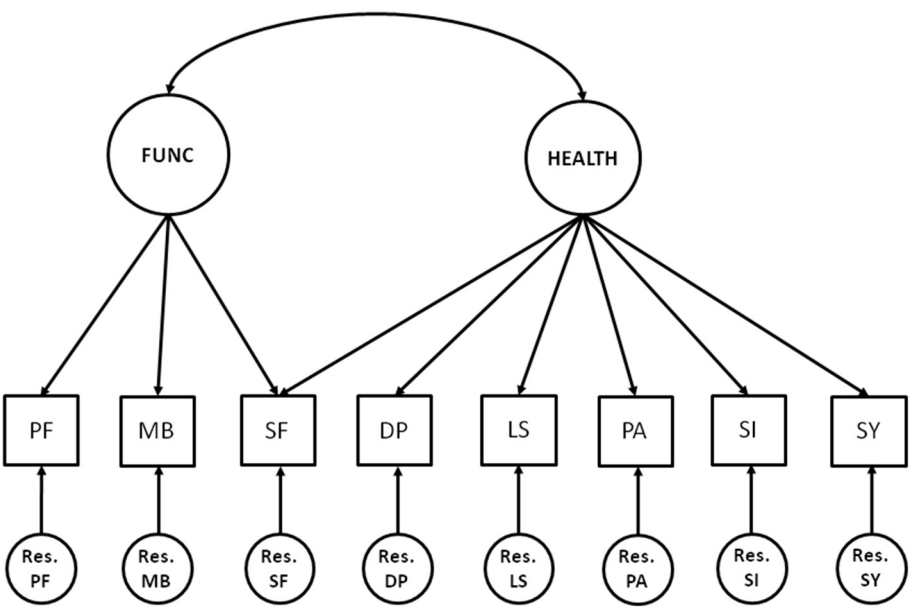


FIGURE 1 | The measurement model. Circles represent latent variables (common and residual factors) and squares represent observed variables (the scale scores). FUNC, functional limitations; HEALTH, health impairments; PF, physical functioning; MB, mobility; SF, social functioning; DP, depression; LS, listlessness; PA, pain; SI, sickness; SY, treatment related symptoms; and Res., Residual factors.

Table 1 | Goodness of overall fit of models in the four-step measurement bias detection procedure.

Model	Description	DF	CHISQ	RMSEA [95% CI]	ECVI [95% CI]
Model 1	Measurement model	4920	9094.7	0.035 [0.034;0.036]	15.59 [15.11; 16.09]
Model 2	Measurement invariance model	5076	9829.9	0.037 [0.036;0.038]	16.13 [15.62; 16.66]
Model 3	Partial measurement invariance model	5040	9318.2	0.035 [0.034;0.037]	15.50 [15.01; 16.01]
Model 4	Curves partial measurement invariance model	5070	9380.8	0.035 [0.034;0.037]	15.49 [15.00; 16.00]

n = 682.

Table 2 | Measurement invariant parameter estimates of the Partial Measurement Invariance Model.

	PF	MB	SF	DP	LS	PA	SI	SY
INTERCEPTS (τ_0)								
	3.03	2.12	2.25	1.98	2.29	**Bias**	**Bias**	1.46
FACTOR LOADINGS (Λ_0)								
FUNC	0.90	0.70	0.29					
HEALTH			0.27	0.39	0.43	0.35	**Bias**	0.19

N = 682; parameter estimates are unstandardized.

was still a significant difference in fit according to the chi-square difference test, comparison of approximate fit using the ECVI difference test indicated that the models can be considered approximately equivalent [$\text{CHISQ}_{diff}(120) = 223.5$, $p < 0.001$; $\text{ECVI}_{diff} = -0.09$]. Therefore, the *Partial Measurement Invariance Model* was retained. All invariant parameters of $\mathbf{\Lambda}_0$ and $\boldsymbol{\tau}_0$, and the measurement bias parameters of the three cases of bias, are given in **Tables 2**,**3**, respectively.

Measurement bias intercept "pain"

The first bias that was detected was a measurement bias of the indicator "pain." The model where the intercept of the indicator "pain" was freely estimated across occasions yielded the largest improvement in model fit [$\text{CHISQ}_{diff}(12) = 287.7$, $p < 0.001$; $\text{ECVI}_{diff} = 0.38$, 95% CI: 0.28–0.49]. Inspection of the measurement bias parameters shows that the estimate of the intercept decreases over the first five measurement occasions and stabilizes around the sixth measurement occasion (see **Table 3**). This indicates that, given equal health impairments, patients report decreasing pain over the first 4 weeks after treatment, after which they report stable pain over time.

To get more insight in the trend of this bias, the measurement bias parameters were modeled as a function of the time of measurement. First, a linear curve was fitted to the bias. This model yielded an intercept and slope parameter that can give insight in the trend of the bias across occasions (see **Figure 2**), but the model did not show a good fit to the data [$\text{CHISQ}_{diff}(11) = 189.9$, $p < 0.001$; $\text{ECVI}_{diff} = 0.24$, 95% CI: 0.16–0.33]. In addition, a selection of non-linear curves was fitted to the measurement bias parameters (see **Figure 2**) of which the quadratic curve showed significant deterioration in fit [$\text{CHISQ}_{diff}(10) = 61.0$, $p < 0.001$; $\text{ECVI}_{diff} = 0.05$, 95% CI: 0.02–0.11], but the inverse curve showed equivalent fit to the model with free intercepts [$\text{CHISQ}_{diff}(10) = 18.7, p = 0.044$; $\text{ECVI}_{diff} = -0.01$]. The slope parameter gives an indication of the steepness and direction of the measurement bias for the first five measurement occasions.

Measurement bias intercept "sickness"

The second step of the measurement bias detection procedure showed that the equality constraint on the intercept of the indicator "sickness" across occasions was not tenable [$\text{CHISQ}_{diff}(12) = 141.9$, $p < 0.001$; $\text{ECVI}_{diff} = 0.17$, 95% CI: 0.10–0.25]. Inspection of the measurement bias parameters shows that the intercept of the indicator "sickness" increases over the first four measurement occasions, after which it decreases and stabilizes around the seventh measurement (see **Table 3**). Thus, given equal health impairments, patients report more sickness in the first 3 weeks after treatment, then report less sickness, and after the sixth week after treatment report a stable, above baseline level of sickness.

A model with a linear curve was fitted to the data, which yielded a non-significant slope parameter estimate (see **Figure 3**), and showed significant deterioration in fit compared to the model with free intercepts [$\text{CHISQ}_{diff}(11) = 138.2$, $p < 0.001$; $\text{ECVI}_{diff} = 0.16$, 95% CI: 0.10–0.25]. As it can be seen from the data that different parts of the trajectory of the intercept follow different trends (i.e., first an increase and then a decrease across measurement occasions), we modeled these trajectories in the bias using piece-wise curves. Piece-wise curves were modeled using additional time coding that applied to only part of the trajectory. In this example, linear piece-wise curves were fitted to the measurement bias parameters of "sickness" (see **Figure 3**), where the model with two piece-wise curves did not show a good fit to the data [$\text{CHISQ}_{diff}(10) = 64.7$, $p < 0.001$; $\text{ECVI}_{diff} = 0.06$, 95% CI: 0.02–0.12], but the model with three piece-wise curves showed equivalent fit to the model with free intercepts [$\text{CHISQ}_{diff}(10) = 11.0, p = 0.274$; $\text{ECVI}_{diff} = -0.02$]. The slope parameters give an indication of the steepness and direction of the measurement bias for the first three measurement occasions, and the deviations from this trend for the fourth to sixth measurement occasions, and the seventh to thirteenth measurement occasions (see **Figure 3**).

Measurement bias factor loading "sickness"

The third bias that was detected was a measurement bias of the indicator "sickness," as freeing the equality constraint on the factor loading across occasions yielded the largest improvement in model fit [$\text{CHISQ}_{diff}(12) = 82.0$, $p < 0.001$; $\text{ECVI}_{diff} = 0.08$, 95% CI: 0.03–0.14]. Inspection of the measurement bias parameters shows that the factor loading increases over the first four measurement occasions, after which it decreases again toward

Table 3 | Measurement bias parameter estimates of the Partial Measurement Invariance Model.

T0	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
INTERCEPT "PAIN"												
2.56	2.41	2.33	2.27	2.22	2.21	2.21	2.21	2.21	2.21	2.21	2.21	2.21
INTERCEPT "SICKNESS"												
1.37	1.49	1.56	1.56	1.52	1.49	1.47	1.46	1.46	1.46	1.45	1.43	1.44
FACTOR LOADING "SICKNESS"												
0.28	0.35	0.40	0.41	0.37	0.34	0.34	0.33	0.33	0.33	0.32	0.29	0.31

N = 682; parameter estimates are unstandardized.

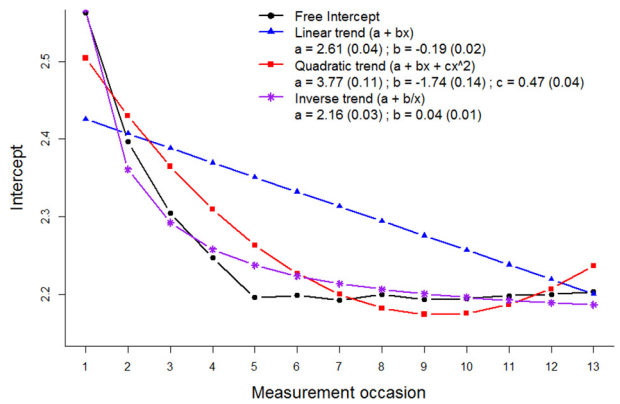


FIGURE 2 | Curves fitted to the measurement bias parameters of the intercept "pain." The black line (circles) represents measurement bias parameter estimates when they are freely estimated across occasions, the blue line (triangles) when they are modeled using a linear curve, the red line (squares) when they are modeled using a quadratic curve, and the purple line (stars) when they are modeled using an inverse curve.

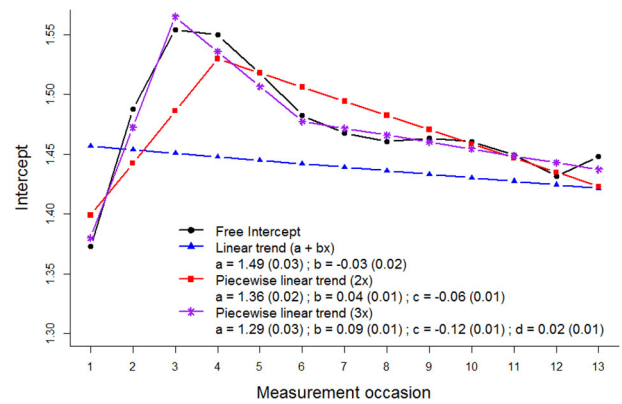


FIGURE 3 | Curves fitted to the measurement bias parameters of the intercept "sickness." The black line (circles) represents measurement bias parameter estimates when they are freely estimated across occasions, the blue line (triangles) when they are modeled using a linear curve, the red line (squares) when they are modeled using two piece-wise linear curves, and the purple line (stars) when they are modeled using three piece-wise linear curves.

baseline level, although it shows a somewhat fluctuating pattern (see **Table 3**). Thus, sickness becomes more important for patients' health impairments in the first 3 weeks after treatment, but then its importance decreases again toward baseline level.

This occurrence of measurement bias was modeled using a linear curve and a piece-wise linear curve (see **Figure 4**). The model with the linear curve showed significant deterioration in fit [$CHISQ_{diff}(11) = 69.7$, $p < 0.001$; $ECVI_{diff} = 0.06$, 95% CI: 0.02–0.12], but the model with two piece-wise curves showed equivalent fit to the model with free factor loadings [$CHISQ_{diff}(10) = 31.1$, $p < 0.001$; $ECVI_{diff} = 0.01$, 95% CI: −0.01–0.05]. The slope parameters give an indication of the steepness and direction of the measurement bias for the first three measurement occasions, and the deviations from this trend for the fourth to thirteenth measurement occasions (see **Figure 4**).

CURVES PARTIAL MEASUREMENT INVARIANCE MODEL

The final model, the *Curves Partial Measurement Invariance Model*, includes the three curves described above to model the measurement biases that were detected. The overall fit of the model was good (RMSEA = 0.035, see **Table 1**) and showed equivalent model fit when compared to the model with no curves fitted to the measurement biases [$CHISQ_{diff}(30) = 62.5$, $p < 0.001$; $ECVI_{diff} = -0.01$].

ASSESSMENT OF CHANGE

The trajectory of the common factor functional limitations (see **Figure 5**) indicates that patients showed a more or less constant trajectory [$CHISQ_{diff}(12) = 39.8$, $p < 0.001$; $ECVI_{diff} = 0.02$, 95% CI: −0.01–0.06]. As the biases that were detected concern the measurement of health impairments, taking into account measurement bias did not affect the trajectory of functional limitations.

The trajectory of health impairments (see **Figure 6**) shows that patients significantly improved [$CHISQ_{diff}(12) = 51.5$, $p < 0.001$; $ECVI_{diff} = 0.03$, 95% CI: 0.001–0.085], although it seems that patients slightly deteriorated again in the last 3 weeks of measurement. Taking into account the measurement biases of the indicators of health impairments affected the trajectory, as it can be seen that health impairments would be generally underestimated across occasions.

DISCUSSION

Measurement invariance is a prerequisite for a valid assessment of change. Longitudinal measurement invariance is usually investigated with a LFM. However, in the situation when there are many measurement occasions the LFM can become of unmanageable size. One solution to this problem is the imposition of

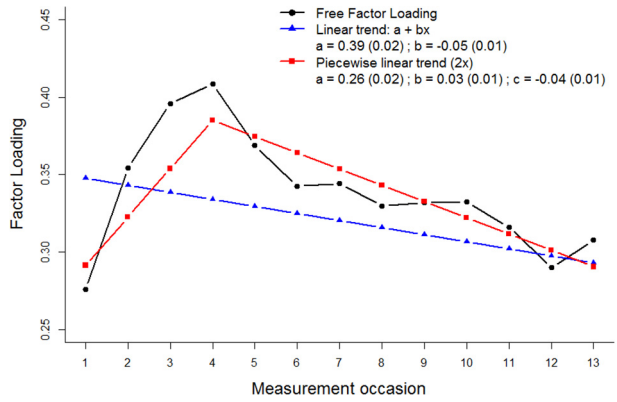


FIGURE 4 | Linear curve of measurement bias parameters of the factor loading "sickness." The black line (circles) represents measurement bias parameter estimates when they are freely estimated across occasions, the blue line (triangles) when they are modeled using a linear curve, and the red line (squares) represents measurement bias parameter estimates when they are modeled using two piece-wise linear curves.

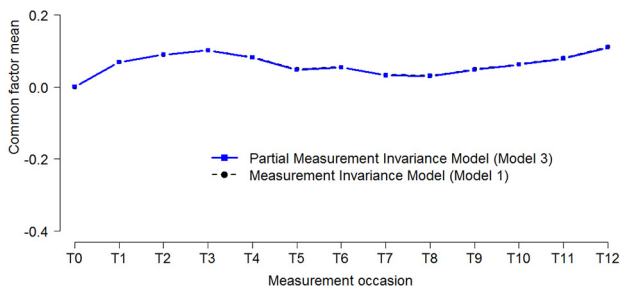


FIGURE 5 | Latent trajectories of functional limitations before and after accounting for measurement bias. The dotted black line (circles) represents estimates of the *Measurement Invariance Model*, and the solid blue (squares) line represents parameter estimates of the *Partial Measurement Invariance Model*, where all measurement biases are incorporated in the model.

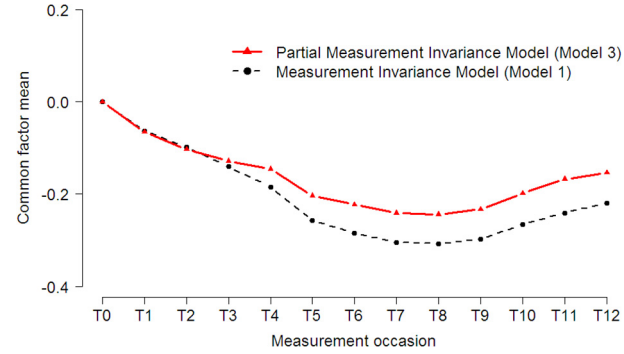


FIGURE 6 | Latent trajectories of health impairments before and after accounting for measurement bias. The dotted black line (circles) represents estimates of the *Measurement Invariance Model*, and the solid red (triangles) line represents parameter estimates of the *Partial Measurement Invariance Model*, where all measurement biases are incorporated in the model.

Kronecker product restrictions to model the multivariate longitudinal structure of the data. In these models Kronecker product restrictions also imply measurement invariance across measurement occasions. As a result, measurement bias across occasion cannot be investigated in the usual way, by testing equality constraints on individual parameters (intercepts and factor loadings). Therefore, to investigate which measurement parameters show violations of measurement invariance (i.e., measurement bias) in Kronecker product restricted models, we propose a modeling procedure that uses additional matrices to accommodate possible bias. This enables the investigation of measurement bias, to account for apparent bias, and use partial measurement invariance to investigate change in common factor means.

The procedure that we propose enables the investigation of measurement invariance in Kronecker product restricted models for multivariate longitudinal data when the number of measurement occasions is large. The procedure of measurement invariance investigation is not different from the usual procedure, but requires alternative modeling as the usual LFM cannot be applied in the situation when invariance restrictions on factor loadings and intercepts are imposed using the Kronecker product. Moreover, with additional matrices that are used to accommodate possible violations of measurement invariance, it is possible to further investigate and model detected biases. This paper therefore contributes to the existing literature on measurement bias detection using structural equation modeling by: (1) using the imposition of Kronecker product restrictions to enable factor analyses of data from a large number of measurement occasions, (2) describing a procedure that enables measurement invariance investigation with Kronecker product restricted models, and (3) modeling the measurement bias parameters to facilitate interpretation of detected biases.

In case of bias, the detected measurement bias can be modeled as a function of the time of measurement using linear or non-linear curves. It should be noted that this technique was used in an exploratory way, e.g., the curve that was fitted to the bias was chosen after inspection of the trajectory of the measurement bias parameters. Interpretation of bias is then facilitated by decreasing the number of parameters to be interpreted, i.e., a slope parameter indicates direction and strength of the trend of the bias across time. Moreover, additional information could be used to test specific hypotheses, for example by incorporating the time of an event (e.g., start of treatment) in modeling the curves.

In our illustrative sample of bone metastases patients imposition of Kronecker product restrictions enabled the analyses of multivariate data from 13 measurement occasions, and the proposed procedure for the investigation of measurement invariance enabled the detection of measurement bias, to account for apparent bias, and use partial measurement invariance to investigate change in HRQL. We found that patients showed a constant trajectory of functional limitations and an improvement of health impairments over time. If measurement bias had not been taken into account, patient's health impairments would generally be underestimated. Moreover, measurement bias was detected in

the intercept of the indicator pain, and in both the intercept and factor loading of the indicator sickness. Given equal health impairments, patients reported decreasing pain over the first 4 weeks after treatment, after which they reported stable pain over time. In addition, given equal health impairments patients reported more sickness in the first 3 weeks after treatment, after which they again reported less sickness. Similarly, the importance of sickness became more important for patients' health impairments in the first 3 weeks and then decreased again toward baseline level. A possible explanation for the bias in pain as a measurement of health impairments could be that the radiotherapy treatment led to a larger decrease in pain than in the other indicators of health impairments. In the measurement of health impairments, patients' reporting of pain would then decrease relative to the other indicators. A possible explanation for the biases in sickness could be that patients experienced side-effects from radiotherapy and that symptoms related to sickness were relatively more prevalent than the other symptoms. Sickness could therefore have become more important to the measurement of health impairments, relative to the other symptoms. As these side-effects usually disappear after a few weeks, this could explain the subsequent decrease in both the reporting of sickness relative to the other symptoms and its importance in the measurement of health impairments. These occurrences of measurement bias and their impact on the assessment of change emphasize the importance of investigating measurement invariance when analyzing longitudinal data. Our proposed procedure enables the investigation of measurement invariance in Kronecker product restricted models, and therefore allows for a more complete interpretation of findings from multivariate longitudinal data.

PRACTICAL GUIDELINES

The introduction of parameter matrices that can accommodate possible violations of measurement invariance enables the investigation of bias in individual factor loading and intercepts. Further investigation of cases of bias is possible through modeling the measurement bias using linear and non-linear curves. The proposed methods not only enable the investigation of measurement bias with longitudinal three-mode models, but can also enhance our understanding of occurrences of measurement bias in multivariate longitudinal data.

ACKNOWLEDGMENTS

This research was supported by the Dutch Cancer Society (KWF grant 2011-4985). Both authors participate in the Research Priority Area Yield of the University of Amsterdam. We would like to thank Y. M. van der Linden for making the data from the Dutch Bone Metastasis Study available for secondary analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.01022/abstract

REFERENCES

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2011). OpenMx: an open source extended structural equation modeling framework. *Psychometrika* 76, 306–317. doi: 10.1007/s11336-010-9200-6

Browne, M. W., and Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behav. Res.* 24, 445–455. doi: 10.1207/s15327906mbr2404_4

Browne, M. W., and Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociol. Methods Res.* 21, 230–258. doi: 10.1177/0049124192021002005

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

Dudgeon, P. (2003). *NIESEM: A Computer Program for Calculating Noncentral Interval Estimates (and Power Analysis) for Structural Equation Modeling* [Computer Software]. Melbourne: University of Melbourne, Department of Psychology.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Oort, F. J. (2001). Three-mode models for multivariate longitudinal data. *Br. J. Math. Stat. Psychol.* 54, 49–78. doi: 10.1348/000711001159429

Sprangers, M. A. G., and Schwartz, C. E. (1999). Integrating response shift into health-related quality of life research: a theoretical model. *Soc. Sci. Med.* 48, 1507–1515. doi: 10.1016/S0277-9536(99)00045-3

Steenland, E., Leer, J., van Houwelingen, H., Post, W. J., van den Hout, W. B., Kievit, J., et al. (1999). The effect of a single fraction compared to multiple fractions on painful bone metastases: a global analysis of the Dutch Bone Metastasis Study. *Radiother.Oncol.* 52, 101–109. doi: 10.1016/S0167-8140(99)00110-3

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behav. Res.* 25, 173–180. doi: 10.1207/s15327906mbr2502_4

Steiger, J. H., and Lind, J. C. (1980). "Statistically based tests for the number of common factors," in *Paper presented at the Annual Meeting of the Psychometric Society* (Iowa City, IA).

Van der Linden, Y. M., Lok, J. J., Steenland, E., Martijn, H., van Houwelingen, H., Marijnen, C. A. M., et al. (2004). Single fraction radiotherapy is efficacious: a further analysis of the Dutch bone metastasis study controlling for the influence of retreatment. *Int. J. Radiat. Oncol. Biol. Phys.* 59, 528–537. doi: 10.1016/j.ijrobp.2003.10.006

World Health Organization. (2002). *Towards a Common Language for Functioning, Disability and Health: the International Classification of Functioning, Disability and Health (ICF)*. Geneva: World Health Organization.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 10 April 2014; paper pending published: 11 May 2014; accepted: 27 August 2014; published online: 23 September 2014.

*Citation: Verdam MGE and Oort FJ (2014) Measurement bias detection with Kronecker product restricted models for multivariate longitudinal data: an illustration with health-related quality of life data from thirteen measurement occasions. Front. Psychol. **5**:1022. doi: 10.3389/fpsyg.2014.01022*

This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in PSYCHOLOGY

ORIGINAL RESEARCH ARTICLE
published: 30 May 2014
doi: 10.3389/fpsyg.2014.00438

Score-based tests of measurement invariance: use in practice

Ting Wang[1], Edgar C. Merkle[1]* and Achim Zeileis[2]

[1] *Department of Psychological Sciences, University of Missouri, Columbia, MO, USA*
[2] *Department of Statistics, Faculty of Economics and Statistics, Universität Innsbruck, Innsbruck, Austria*

Edited by:
Alain De Beuckelaer, Radboud University Nijmegen, Netherlands

Reviewed by:
Theo K. Dijkstra, University of Groningen, Netherlands
Conor Vivian Dolan, Vrije Universiteit Amsterdam, Netherlands

****Correspondence:***
Edgar C. Merkle, Department of Psychological Sciences, University of Missouri, 28A McAlester Hall, Columbia, MO 65211, USA
e-mail: merklee@missouri.edu

In this paper, we consider a family of recently-proposed measurement invariance tests that are based on the *scores* of a fitted model. This family can be used to test for measurement invariance w.r.t. a continuous auxiliary variable, without pre-specification of subgroups. Moreover, the family can be used when one wishes to test for measurement invariance w.r.t. an ordinal auxiliary variable, yielding test statistics that are sensitive to violations that are monotonically related to the ordinal variable (and less sensitive to non-monotonic violations). The paper is specifically aimed at potential users of the tests who may wish to know (1) how the tests can be employed for their data, and (2) whether the tests can accurately identify specific models parameters that violate measurement invariance (possibly in the presence of model misspecification). After providing an overview of the tests, we illustrate their general use via the R packages *lavaan* and *strucchange*. We then describe two novel simulations that provide evidence of the tests' practical abilities. As a whole, the paper provides researchers with the tools and knowledge needed to apply these tests to general measurement invariance scenarios.

Keywords: measurement invariance, factor analysis, lavaan, parameter stability, ordinal variable, structural equation modeling

1. INTRODUCTION

Some of the papers in this special issue focus on the topic of approximate measurement invariance: we know that strict hypotheses of measurement invariance do not hold exactly across different groups, and this should be reflected in corresponding tests of measurement invariance. Under a Bayesian approach, we may implement the idea of approximate invariance (e.g., Muthén and Asparouhov, 2013) by replacing across-group equality constraints on parameters with informative prior distributions. In this paper, we describe an alternative approach: the development of test statistics that are especially sensitive to violations that are monotonic w.r.t. the variable of interest (and less sensitive to violations that are non-monotonic w.r.t. the variable of interest). Because monotonic violations are more likely to be interpretable and interesting to the researcher, we can gain more power to detect these violations by de-emphasizing other types of violations. The resulting test statistics are specifically applicable to situations where one wishes to test for measurement invariance with respect to an ordinal variable, and they are special cases of a family of tests that may be used to study measurement invariance w.r.t. continuous, categorical, and ordinal variables.

The study of measurement invariance w.r.t. categorical auxiliary variables (via, e.g., likelihood ratio tests) is popular and well known, and ordinal auxiliary variables are typically treated as categorical in measurement invariance contexts. The study of measurement invariance w.r.t. continuous variables is newer and less known: along with the family described here, other methods include moderated factor models (Purcell, 2002; Bauer and Hussong, 2009; Molenaar et al., 2010) and factor mixture models (Dolan and van der Maas, 1998; Lubke and Muthén, 2005). These methods require estimation of a model of greater complexity, while the tests described in this paper work solely on the output of a traditional factor model (see Merkle and Zeileis, 2013, for further comparison of these methods). These methods all assume that the estimated model is correctly specified, save possibly for differences in parameter values between individuals.

The family of tests described here have recently been applied to the study of measurement invariance (Merkle and Zeileis, 2013; Merkle et al., 2014), though their practical application has been limited to a small set of simulations and data examples. In this paper, we provide a detailed illustration of the tests' use and performance under scenarios likely to be encountered in practice. While the previous papers have described and studied the tests under ideal conditions, we focus here on two topics of interest to the applied researcher: software considerations for carrying out the tests, and test performance under model misspecification. The latter issue is particularly important because, in practice, all models are misspecified. Hence, practically-useful tests of measurement invariance should be robust to model misspecification.

In the following sections, we first briefly review the theoretical framework of the proposed tests and provide a short tutorial illustrating the use of the tests in R (R Core Team, 2013). Next, we study the tests' performance in simulations that mimic practical research scenarios. Finally, we provide some further discussion on the tests' use in practice.

2. BACKGROUND

This section contains background and discussion of the proposed statistics as applied to structural equation models (SEMs); for a more detailed account, see Merkle and Zeileis (2013) and Merkle

et al. (2014). For details on the statistics' application to general statistical models, see Zeileis and Hornik (2007).

As currently implemented for SEM, the statistical tests described in this paper can be applied to models that are estimated via a multivariate normal or Wishart likelihood (or discrepancy) function, with extension to other discrepancy functions being conceptually straightforward. The tests are carried out following model estimation, making use of output associated with the fitted model. In general, we fit a model that restricts parameters to be equal across observations, then carry out a *post hoc* test to examine whether specific parameters vary across observations. This procedure is similar in spirit to the calculation of modification indices (Bentler, 1990) and to Lagrange multiplier tests (Satorra, 1989), and, in fact, those statistics can be viewed as special cases of the family described here.

Following model estimation, the tests primarily work on the *scores* of the fitted model; these are defined as

$$s(\theta; x_i) = \left(\frac{\partial \ell(\theta; x_i)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta; x_i)}{\partial \theta_k} \right)^\top, \; i = 1, \ldots, n, \qquad (1)$$

where $\ell(\theta; x_i)$ is the likelihood associated with individual i and θ is a k-dimensional parameter vector. The corresponding maximum likelihood estimate $\hat{\theta}$ solves the first order condition: $\sum_{i=1}^{n} s(\hat{\theta}; x_i) = 0$.

To verbally describe Equation (1), each individual has k scores describing the extent to which the fitted model describes that particular individual. These scores are similar to residuals and, in fact, the tests can be applied directly to residuals in other contexts (see Zeileis and Hornik, 2007): we can roughly interpret scores near zero as providing a "good" description of an individual, with scores far from zero providing a "bad" description of an individual. This is only a rough interpretation as, even when measurement invariance holds, some individuals' scores will be further from zero than others. However, under measurement non-invariance, the scores will differ for different subgroups of individuals (say, scores in subgroup A tend to be negative and scores in subgroup B tend to be positive). Each of the k scores represents one model parameter, which, as further described below, allows us to test subsets of model parameters for invariance. While scores can be obtained under the multivariate normal likelihood (discrepancy) function and alternatives such as generalized least squares, most SEM software fails to supply the scores to the user.

To use the scores for testing, we order individuals according to an auxiliary variable V (the variable against which we are testing measurement invariance) and look for "trends" in the scores. For example, imagine that we are testing for measurement invariance w.r.t. age. If there exists measurement non-invariance w.r.t. age, then some parameter estimates may be too large for young individuals and too small for old individuals (say). This result would be reflected in the scores, where young individuals' scores may be greater than zero and old individuals' scores less than zero (though the sign of the scores will depend on whether a function is being minimized or maximized). Conversely, if measurement invariance holds, then all individuals' scores will fluctuate randomly around zero.

To formalize these ideas, we define a suitably scaled cumulative sum of the ordered scores. This may be written as

$$\boldsymbol{B}(t; \hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{I}}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \mathbf{s}(\hat{\boldsymbol{\theta}}; x_{(i)}) \qquad (0 \leq t \leq 1) \qquad (2)$$

where $\hat{\boldsymbol{I}}$ is an estimate of the information matrix, $\lfloor nt \rfloor$ is the integer part of nt (i.e., a floor operator), and $x_{(i)}$ reflects the individual with the i-th smallest value of the auxiliary variable V. While the above equation is written in general form, we can restrict the value of t in finite samples to the set $\{0, 1/n, 2/n, 3/n, \ldots, n/n\}$. We focus on how the cumulative sum fluctuates as more individuals' scores are added to it, e.g., starting with the youngest and ending with the oldest individual if age is the auxiliary variable of interest. The summation is premultiplied by an estimate of the inverse square root of the information matrix, which serves to decorrelate the fluctuation processes associated with individual model parameters while preserving the behavior of individual parameters' fluctuations.

Under the hypothesis of measurement invariance, a central limit theorem can be used to show that the fluctuation of the above cumulative sum follows a Brownian bridge (Hjort and Koning, 2002). This result allows us to calculate p-values and critical values for test statistics under the hypothesis of measurement invariance. We can obtain test statistics associated with all model parameters and with subsets of model parameters.

Multiple test statistics are available, depending on how one summarizes the behavior of the cumulative sum of scores. For example, one could take the absolute maximum that the cumulative sum attains for any parameter of interest, resulting in a *double max* statistic (the maximum is taken across parameters and individuals). Alternatively, one could sum the (squared) cumulative sum across parameters of interest and take the maximum or the average across individuals, resulting in a *maximum Lagrange multiplier* statistic and *Cramér-von Mises* statistic, respectively (see Merkle and Zeileis, 2013, for further discussion). These statistics are given by

$$DM = \max_{i=1,\ldots,n} \max_{j=1,\ldots,k} |\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}| \qquad (3)$$

$$CvM = n^{-1} \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, \qquad (4)$$

$$\max LM = \max_{i=\underline{i},\ldots,\bar{i}} \left\{ \frac{i}{n} \left(1 - \frac{i}{n} \right) \right\}^{-1} \sum_{j=1,\ldots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2. \qquad (5)$$

Critical values associated with DM can be obtained analytically, while critical values associated with the other statistics can be obtained from direct simulation (Zeileis, 2006) or from more refined techniques (Hansen, 1997). This issue should not be important to the user, as critical values are obtained directly from the R implementation described later.

Importantly, the above statistics were derived for situations where individuals are uniquely ordered according to the auxiliary variable. This is not always the case for measurement invariance

applications, where the auxiliary variable is often ordinal. To remedy this situation, Merkle et al. (2014) extended the framework to situations where one has an ordinal auxiliary variable of interest. Essentially, one allows all individuals with the same value of the auxiliary variable to enter into the cumulative sum at the same time. Analogous test statistics are then computed, with modified critical values being adopted to reflect the change in the statistics' computation.

For an ordinal auxiliary variable with m levels, these modifications are based on t_ℓ ($\ell = 1, \ldots, m-1$), which are the empirical, cumulative proportions of individuals observed at the first $m-1$ levels. The modified statistics are then given by

$$WDM_o = \max_{i \in \{i_1, \ldots, i_{m-1}\}} \left\{ \frac{i}{n}\left(1 - \frac{i}{n}\right) \right\}^{-1/2} \max_{j=1,\ldots,k} |\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}|, \quad (6)$$

$$\max LM_o = \max_{i \in \{i_1, \ldots, i_{m-1}\}} \left\{ \frac{i}{n}\left(1 - \frac{i}{n}\right) \right\}^{-1} \sum_{j=1,\ldots,k} \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{ij}^2, \quad (7)$$

where $i_\ell = \lfloor n \cdot t_\ell \rfloor$ ($\ell = 1, \ldots, m-1$). Critical values associated with the WDM_o statistic can be obtained directly from a multivariate normal distribution (see Hothorn and Zeileis, 2008), while critical values associated with $\max LM_o$ can be obtained via simulation. This simulation is somewhat computationally intensive and, in practice, takes about 10 min on the authors' computers when 50,000 replications are sampled from the approximate asymptotic distribution. However, the wait is often worth it, as Merkle et al. (2014) found the performance of the $\max LM_o$ statistic to have more power than the WDM_o statistic and the traditional likelihood ratio test statistic when the measurement invariance violation is monotonic with the ordinal variable.

Finally, if the auxiliary variable V is only nominal/categorical, the cumulative sums of scores can be used to obtain a Lagrange multiplier statistic. This test statistic can be formally written as

$$LM_{uo} = \sum_{\ell=1,\ldots,m} \sum_{j=1,\ldots,k} \left(\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_\ell j} - \boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_{\ell-1} j} \right)^2, \quad (8)$$

where $\boldsymbol{B}(\hat{\boldsymbol{\theta}})_{i_0 j} = 0$ for all j. This statistic is asymptotically equivalent to the usual, likelihood ratio test statistic, and it is advantageous over the likelihood ratio test because it requires estimation of only one model (the restricted model). We make use of this advantage in the simulations, described later.

3. TUTORIAL

In this section, we demonstrate how the above tests can be carried out in R, using the package *lavaan* (Rosseel, 2012) for model estimation and *strucchange* (Zeileis et al., 2002; Zeileis, 2006) for testing. We use data from Froh et al. (2011) concerning the applicability of adult gratitude scales to youth, available in the R package *psychotools* (Zeileis et al., 2013). The data consist of responses to three adult gratitude scales from $n = 1401$ youth aged 10–19 years. The original authors were specifically interested in whether the scales were measurement invariant across age. Because the sample size at each age was unbalanced, the authors created age groups of approximately equal sample size. In the examples below, we test for measurement invariance across these age groups. For illustrative purposes, we conduct multiple tests and compare them to the traditional significance level of 0.05. In practice, however, one should generally adjust the significance level for the number of tests carried out. Additionally, because measurement invariance researchers often have large sample sizes, cross-validation methods can be useful to help verify the test results.

We focus on measurement invariance of the factor loadings associated with one of the scales in the dataset, the GQ-6 scale (McCullough et al., 2002). This scale consists of five Likert scale items (there is a sixth item that is omitted from analyses, following Froh et al.) with seven points each. We fit a one-factor model to these items, examining whether the factor model parameters are invariant with respect to age group. While the age group variable is best considered ordinal, for demonstration we consider its treatment as categorical, continuous, and ordinal. Each of these treatments is described below in a separate section.

3.1. CATEGORICAL TREATMENT

Measurement invariance is most often tested using multiple groups models (see, e.g., van de Schoot et al., 2012). This amounts to assuming that our auxiliary variable is categorical (i.e., unordered), which is not true for the age groups in the data. However, we demonstrate the procedure for completeness.

To conduct the analysis, we first load the data and keep only complete cases for simplicity (though the tests can be applied to incomplete data).

```
R> data("YouthGratitude", package = "psychotools")
R> compcases <- apply(YouthGratitude[, 4:28], 1,
+                     function(x) all(x %in% 1:9))
R> yg <- YouthGratitude[compcases, ]
```

Next, we fit two models in *lavaan*: a one-factor model where loadings are restricted to be equal across age groups, and a one-factor model where loadings are free across age groups. This allows us to test a hypothesis of weak measurement invariance that was of interest to the original researchers (though, for ordinal variables, all types of measurement invariance can be examined via the tests described previously). By default, the code below sets the scale by fixing the first loading to 1.

```
R> restr <- cfa("f1 =~ gq6_1 + gq6_2 + gq6_3 + gq6_4 +
+                      gq6_5",
+               data = yg, group = "agegroup",
+               meanstructure = TRUE,
+               group.equal = "loadings")
R> full <- cfa("f1 =~ gq6_1 + gq6_2 + gq6_3 + gq6_4 +
+                     gq6_5",
+              data = yg, group = "agegroup",
+              meanstructure = TRUE)
```

Finally, we can get the results of a likelihood ratio test via the `anova()` function, which implies that the GQ-6 violates measurement invariance.

```
R> anova(full, restr)

Chi Square Difference Test
```

```
      Df   AIC   BIC Chisq Chisq diff Df diff Pr(>Chisq)
full  30 18947 19414   139
restr 50 18945 19308   177       38.1      20     0.0087
```

To obtain the asymptotically equivalent LM_{uo} (Equation 8), we can use the `sctest()` function from *strucchange*:

```
R> sctest(restr, order.by = yg$agegroup, parm = 1:4,
+         vcov = "info", functional = "LMuo")

       M-fluctuation test

data:  restr
f(efp) = 31.4, p-value = 0.05018
```

This command specifies that we assess the parameters 1–4 of model `restr` after ordering the observations according to `agegroup`. Additionally, the observed information matrix is used as the variance-covariance matrix. Note that the model parameters 1–4 are the factor loadings supplied by *lavaan*, which can be seen by inspecting `coef(restr)`. This also leads to somewhat smaller test statistics that are very close to being significant at the 5% level.

Because our sample size is large, the likelihood ratio test is known to be sensitive to small measurement invariance violations (Bentler and Bonett, 1980). That is, the LRT and LM_{uo} test from Equation (8) are sensitive to small measurement invariance violations that are not likely to be of interest to researchers. For example, imagine that the 15-year-olds' parameters are slightly different than the other age groups. The 15-year-olds are in the middle of the age groups, and there is not likely to be any theoretical justification for 15-year-olds differing from every other age group. One solution to this problem would be the Bayesian, approximate invariance methods described in the introduction (Muthén and Asparouhov, 2013). Alternatively, we can use the "ordinal" score-based statistics (from Equations (6), (7)) to obtain tests that are sensitive to the ordering of age.

3.2. CONTINUOUS TREATMENT

If we are interested in measurement invariance violations that are monotonic with the age groups, it is perhaps simplest to treat the age groups as continuous. In doing so, we can use the statistics from Equations (3–5). That is, we can fit a model whose parameters are restricted to be equal across all individuals and then examine how individuals' scores $s(\hat{\theta}; x_i)$ fluctuate with their age (where age ties are broken arbitrarily, using the original order of the observations within each age group). This is demonstrated below, with similar code being useful when one is testing for measurement invariance w.r.t. truly continuous variables.

Again, we employ the `sctest()` function to assess parameters 1–4 from the restricted model `restr` after ordering w.r.t. `agegroup`:

```
R> dm <- sctest(restr, order.by = yg$agegroup,
+               parm = 1:4, vcov = "info",
+               functional = "DM")
R> cvm <- sctest(restr, order.by = yg$agegroup,
+                parm = 1:4, vcov = "info",
+                functional = "CvM")
R> maxlm <- sctest(restr, order.by = yg$agegroup,
+                  parm = 1:4, vcov = "info",
+                  functional = "maxLM")
R> c(dm$p.value, cvm$p.value, maxlm$p.value)

[1] 0.03804 0.11557 0.00414
```

We see that two of the three p-values output at the end of the code are larger than that associated with the LRT (with the CvM statistic being non-significant).

The tests carried out here assume a unique ordering of individuals by age, but this is obviously not the case. To compute the statistics and p-values, the *strucchange* package implicitly employed the (arbitrary) ordering of individuals who are tied on age. If we were to change this ordering, the resulting statistics and p-values would also change, potentially switching significant results to being non-significant and vice versa. Clearly, this is problematic. To accurately account for the multiple observations at the same age level, we must use the ordinal tests from Equations (6) and (7). These are described next.

3.3. ORDINAL TREATMENT

The main difference between the ordinal test statistics and their continuous counterparts is that the ordinal statistics are unchanged when re-ordering individuals within the same age group. To compute the test statistics, we allow the scores of all tied individuals to enter the cumulative sum (Equation (2)) simultaneously. This results in modified critical values and test statistics that are sensitive to measurement invariance violations that are monotonic w.r.t. age group.

To carry out the tests, we can rely on the same function that we used for the continuous test statistics. As mentioned previously, calculation of the max LM_o statistic (Equation (7)) can be lengthy from the need to simulate critical values (though see the end of this section, which provides a partial speed-up).

```
R> wdmo <- sctest(restr, order.by = yg$agegroup,
+                 parm = 1:4, vcov = "info",
+                 functional = "WDMo")
R> maxlmo <- sctest(restr, order.by = yg$agegroup,
+                   parm = 1:4, vcov = "info",
+                   functional = "maxLMo")

R> c(wdmo$p.value, maxlmo$p.value)

[1] 0.0588 0.0970
```

In computing the ordinal test statistics, we obtain $p = 0.059$ and $p = 0.097$, respectively.[1] Both p-values are clearly larger than that of the likelihood ratio test and neither is significant at $\alpha = 0.05$. This provides evidence that there is no measurement invariance violation that is monotonic with age group. Instead, given the large sample size, the likelihood ratio test may be overly sensitive to anomalous, non-monotonic violations at one (or a few) age groups.

[1]To replicate both p-values exactly, R's random seed needs to be set by `set.seed(1090)` prior to each `sctest()` call.

In addition to test statistics, "instability plots" can be generated by setting `plot = TRUE` in the `sctest()` calls above. **Figure 1** displays the resulting plots, which represent the ordinal statistics' fluctuations across levels of age group. The *x*-axis reflects age group and the *y*-axis reflects test statistic values (larger values reflect more instability), with the dashed horizontal lines reflecting critical values. The hypothesis of measurement invariance is rejected if the sequence of test statistics crosses the critical value. While the measurement invariance tests are non-significant, the plots imply some instability in the older age groups (15, 16).

Finally, if the user anticipates multiple calculations of the max LM_o statistic for a specific dataset, it is possible to save time by simulating critical values once and re-using them for multiple tests. We can use the `ordL2BB()` function to generate critical values and store them in an object `mLMo`, say. Then, this object can be employed to obtain the test statistic in the usual manner.

```
R> mLMo <- ordL2BB(yg$agegroup)
R> maxlmo <- sctest(restr, order.by = yg$agegroup,
+                   parm = 1:4, vcov = "info",
+                   functional = mLMo)
```

The `ordL2BB()` command automatically generates critical values for testing 1–20 parameters at a time. If only a smaller number of parameters (e.g., only up to 6) is to be tested, some computation time can be saved by setting the `nproc` argument accordingly (e.g., `nproc = 1:6`). In the same way, `nproc` can be employed to simulate higher-dimensional fluctuation processes suitable for testing more parameters. One can re-use `mLMo` in this manner for further tests of the youth gratitude data. Critical values must be resimulated for new data, however, because they depend on the proportion of individuals observed at each level of the ordinal variable (denoted t_ℓ for Equation (7)).

In the above sections, we have illustrated the score-based tests' computation in R. We suspect that the ordinal tests will be most popular with users, because measurement invariance tests are typically carried out across categories (ordered or not), as opposed to continuous variables. Thus, in the sections below, we conduct novel simulations to study the ordinal statistics' expected behavior in practice. In particular, we wish to study (1) the extent to which the ordinal statistics attribute measurement invariance violations to the correct parameter(s), and (2) the extent to which the tests are robust to model misspecification. These issues are especially important to examine because SEMs are typically complex, with many inter-related parameters that may exhibit measurement invariance. Previous applications of score-based tests have typically focused on regression-like models with only a small number of parameters that may exhibit instability (e.g., Zeileis and Hornik, 2007). Thus, the simulations here provide general evidence about the extent to which the tests accurately capture instabilities in complex models.

4. SIMULATION 1

In Simulation 1, we examined the extent to which the proposed tests can "localize" a measurement invariance violation. If, say, a factor loading violates measurement invariance, it is plausible that this violation impacts other parameter estimates, including factor covariances, intercepts or the unique variance associated with the manifest variable in question. Thus, the goal of the Simulation 1 is to examine the extent to which the proposed tests attribute the measurement invariance violation to the parameters that are truly in violation.

4.1. METHODS

To examine these issues, we generated data from a two-factor model with three indicators each (see **Figure 2**). The measurement invariance violation occurred in one of four places: the factor loading associated with Scale 1 (λ_{11}), the intercept (μ_{11}), the unique variance (ψ_{11}), or the factor covariance (ϕ_{12}). Note that the latter violation is not necessarily a measurement invariance (e.g., Meredith, 1993), but it is still a parameter instability that can occur in this type of model. We then tested for measurement invariance (parameter instability) in seven subsets of parameters: each of the four individual parameters noted above, all six factor loadings, all six unique variances and all six intercepts.

Power and Type I error were examined across three sample sizes ($n = 120, 480, 960$), three numbers of categories ($m =$

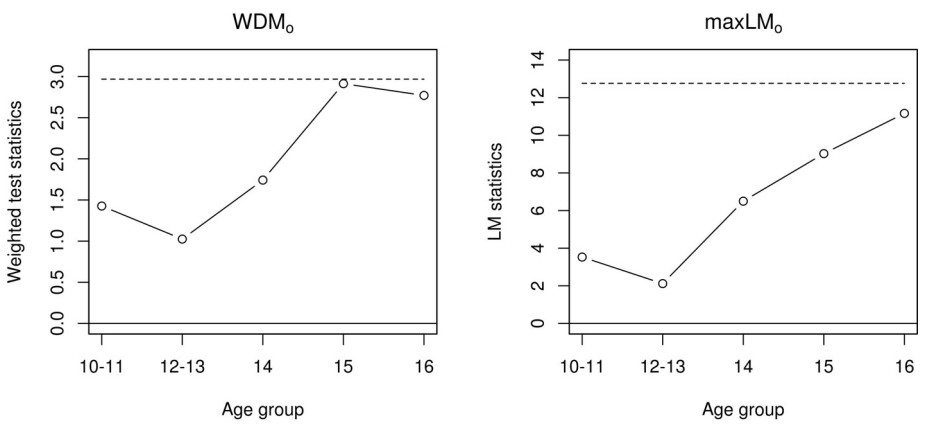


FIGURE 1 | Fluctuation processes for the *WDM$_o$* statistic (left panel) and the max *LM$_o$* statistic (right panel).

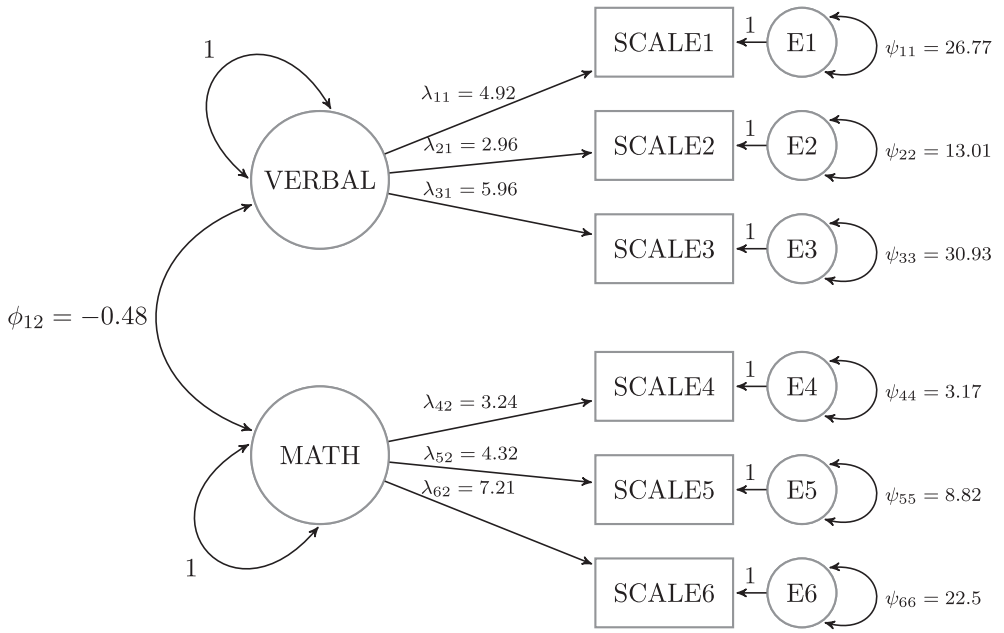


FIGURE 2 | General model used for the simulations.

4, 8, 12), and 17 magnitudes of invariance violations (described in the following sentences). The measurement invariance violations began at level $1 + m/2$ of the auxiliary variable V and were consistent thereafter: individuals below level $1 + m/2$ of V deviated from individuals at or above level $1 + m/2$ by d times the parameters' asymptotic standard errors (scaled by $\sqrt{n}$), with $d = 0, 0.25, 0.5, \ldots, 4$ (see replication code for specific values of the standard errors). For each combination of sample size (n) × violation magnitude (d) × violating parameter × categories (m), 5000 datasets were generated and tested. Statistics from Equations (6–8) were examined. As mentioned previously, Equation (8) is asymptotically equivalent to the usual likelihood ratio test. Thus, this statistic provides information about the relative performance of the ordinal statistics vs. the LRT.

In all conditions, we maintained equal sample sizes in each subgroup of the ordinal variable. Aside from the parameter changes that reflect measurement invariance, the fitted models matched the data generating model.

4.2. RESULTS

Full simulation results are presented in **Figures 3–6**. **Figure 3** displays power curves as a function of violation magnitude in the factor loading λ_{11}, with the parameters being tested changing across rows, the number of levels m of the ordinal variable V across columns, and lines reflecting different test statistics. **Figures 4–6** display similar power curves when the factor covariance ϕ_{12}, error variance ϵ_{11}, and intercept μ_{11} violate measurement invariance, respectively. In these figures, we generally show tests associated with parameters that exhibited non-zero power curves. For example, in **Figure 3**, the middle row shows that power for tests of ψ_{11} stays near zero for all values of m and d. Similar rows have been omitted from this figure and other figures.

Within each panel of **Figures 3–6**, the three lines reflect the three test statistics. It is seen that the two ordinal statistics exhibit similar results, with max LM_{uo} demonstrating lower power across all situations. This demonstrates the sensitivity of the ordinal statistics to invariance violations that are monotonic with V. In situations where only one parameter is tested, WDM_o and max LM_o exhibit equivalent power curves. This is because, when only one parameter is tested, the statistics are equivalent.

From these figures, one generally observes that the tests isolate the parameter violating measurement invariance. Additionally, the tests have somewhat higher power to detect measurement invariance violations in the factor loading, factor covariance, and intercept parameters, as opposed to the error variance parameter. Finally, simultaneous tests of all factor loadings, all intercepts, or all error parameters result in decreased power, as compared to the situation where one tests only the violating parameter. This occurs because, in testing a subset of parameters (only one of which violates measurement invariance), we are dampening the signal of a measurement invariance violation. This "dampening" effect is more apparent for the max LM_o statistic, because it involves a sum across all tested parameters (see Equation 7). Conversely, WDM_o takes the maximum over parameters (Equation 6), so that invariant parameters have no impact on this statistic.

In summary, we found that the proposed tests can attribute measurement invariance violations to the correct parameter. This provides evidence that, in practice, one can have confidence in the tests' abilities to locate the measurement invariance violation. Of course, this statement is qualified by the fact that, in this simulation, the model was correctly specified. In the following simulation, we examine the tests' performance in the likely situation of model misspecification.

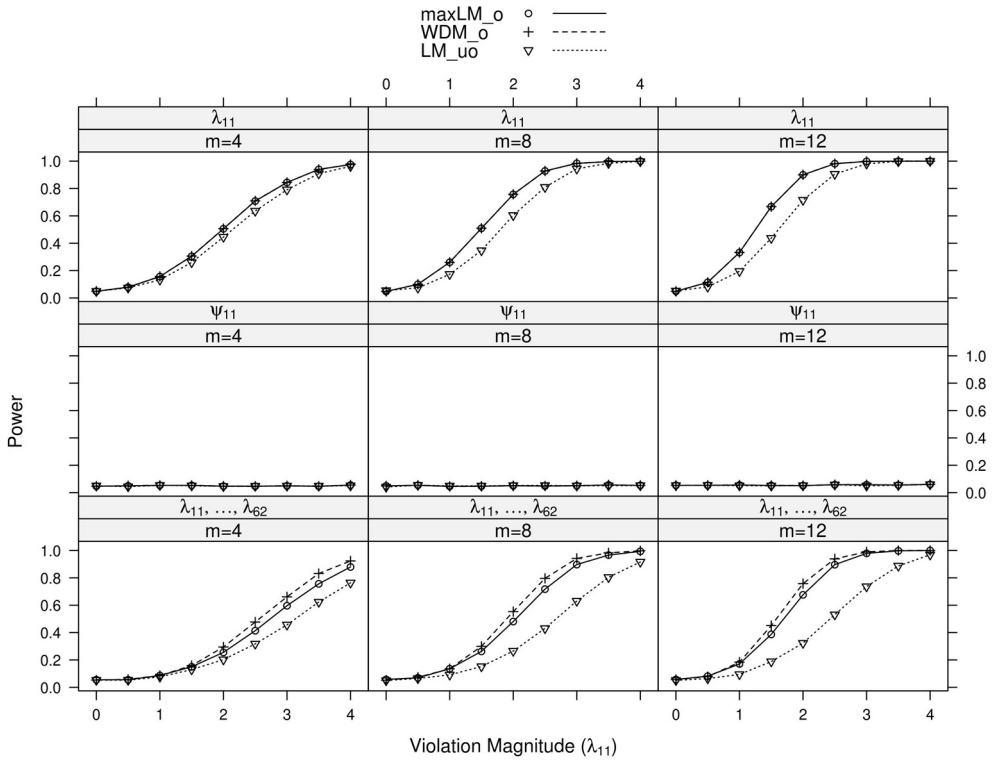


FIGURE 3 | Simulated power curves for max LM_o, WDM_o, and LM_{uo} across three levels of the ordinal variable *m* and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 1. The parameter violating measurement invariance is λ_{11}. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable *m*.

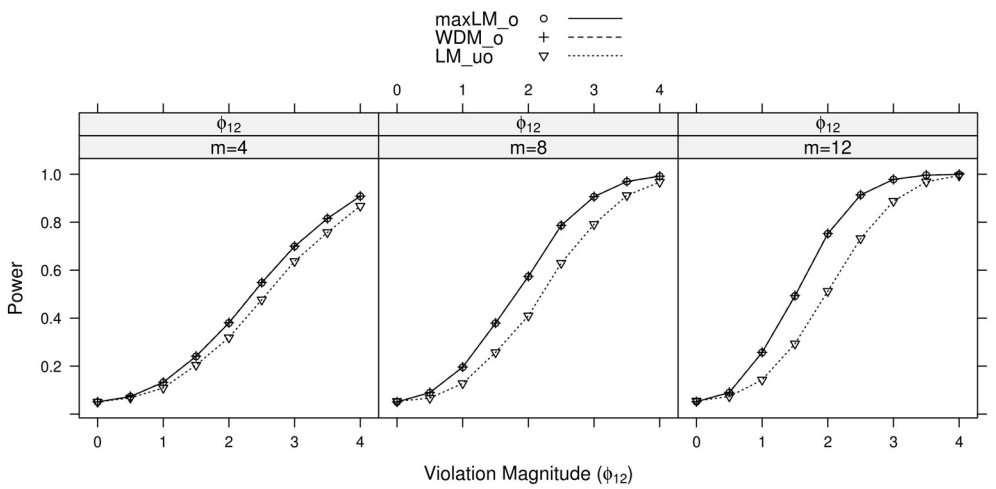


FIGURE 4 | Simulated power curves for max LM_o, WDM_o, and LM_{uo} across three levels of the ordinal variable *m*, and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 1. The parameter violating measurement invariance is ϕ_{12}. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable *m*.

5. SIMULATION 2

In Simulation 2, we examine the extent to which the results of Simulation 1 are robust to model misspecification. Specifically, we generate data from the factor analysis model used in the previous section, except that the model contains an extra loading from the second factor to Scale 1. The estimated model matches that displayed in **Figure 2**, however, resulting in model misspecification. The goal of this simulation is to examine the proposed

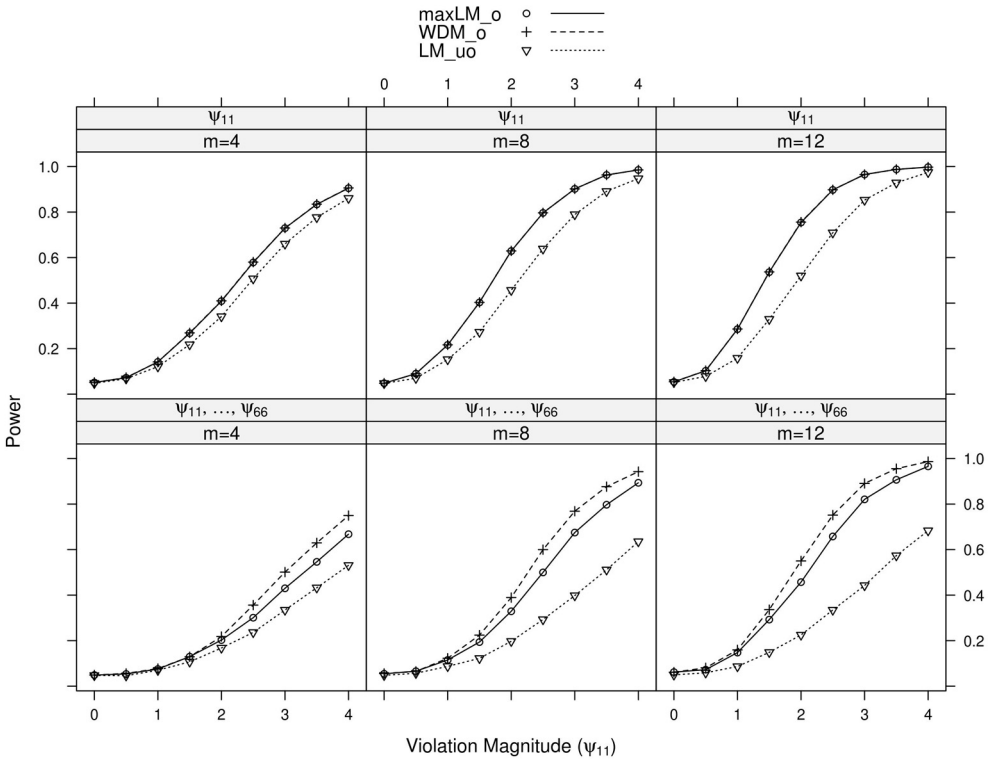


FIGURE 5 | Simulated power curves for max LM_o, WDM_o, and LM_{uo} across three levels of the ordinal variable *m* and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 1. The parameter violating measurement invariance is ψ_{11}. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable *m*.

statistics' power to detect measurement invariance violations (and to attribute the violation to the correct parameter) under this misspecification.

5.1. METHOD

A measurement invariance violation could occur in each of the four parameters from Simulation 1 (factor loading, factor covariance, unique variance, and intercept), and a violation could also occur in the extra, unmodeled loading. In each condition, a single parameter exhibited the violation. Sample size and magnitude of measurement invariance violation were manipulated in the same way as they were in Simulation 1. The tested parameters were also the same as Simulation 1.

5.2. RESULTS

Results of primary interest are conditions where the unmodeled loading violates measurement invariance. A subset of results is displayed in **Figure 7**. One can generally observe that tests of the first loading and unique variance exhibited high "power," which is actually a high Type I error rate here. This Type I error is also observed when testing all loadings and all unique variances (see the Supplementary Material). Tests associated with the factor covariance and intercept did not demonstrate this error, however. In terms of specific statistic performance, max LM_o and WDM_o demonstrated higher Type I error than LM_{uo} in each panel, especially with increasing levels. This is likely because the unmodeled loading's non-invariance was monotonic with V; if it were not monotonic, we would expect LM_{uo} to have higher Type I error.

When the parameter violating measurement invariance was modeled, results were generally the same as Simulation 1. When the modeled factor loading, λ_{11}, violated measurement invariance, the statistics were generally able to pick up the violation despite the misspecification. Similar results were observed when the unique variance, intercept and factor covariance parameters violated measurement invariance; these results are all shown in the Supplementary Material. In particular, power of the ordered statistics was higher than power of the unordered statistic in each panel.

In summary, the proposed test statistics appear robust to unmodeled loading parameters, when the unmodeled loading does not violate measurement invariance and when the rest of the model is correctly specified (save for the measurement invariance violation). If the unmodeled loading does violate measurement invariance, the tests can still detect measurement invariance violations. The violations are assigned to modeled parameters that do not violate measurement invariance, however. The impacted parameters include the error variance and other loadings associated with the manifest variable that has an unmodeled loading. Thus, as for other tests of measurement invariance, it is important to study the extent to which the hypothesized model includes all parameters of importance (i.e., the extent to which the model is well specified).

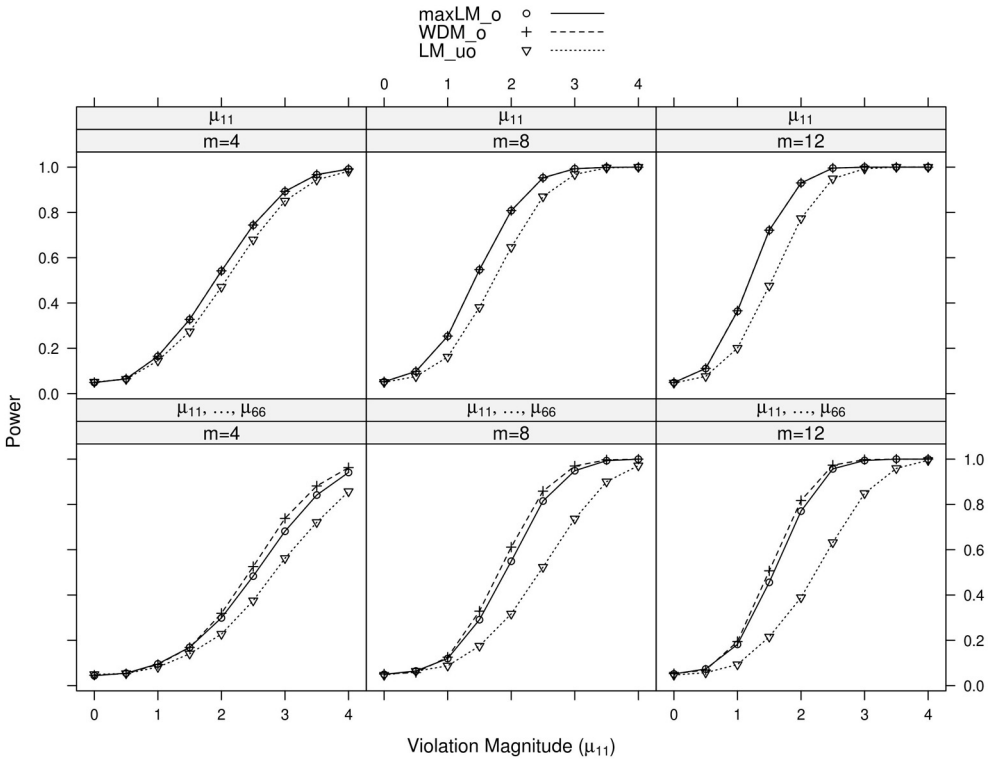


FIGURE 6 | Simulated power curves for max LM_o, WDM_o, and LM_{uo} across three levels of the ordinal variable *m* and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 1. The parameter violating measurement invariance is μ_{11}. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable *m*.

One could begin to study model misspecification by fitting models with different discrepancy functions (say, a multivariate normal function and a generalized least squares function). If parameter estimates differ greatly across the functions, then this implies model misspecification. Additionally, if one has a large sample size, one could split the data into subgroups and examine consistency of results across subgroups. These issues are important for all the tests discussed here (score-based or otherwise).

6. GENERAL DISCUSSION

In this paper, we first described a novel family of test statistics for measurement invariance and illustrated their use via the R packages *lavaan* and *strucchange*. Next, we examined these statistics' abilities to identify the parameter violating measurement invariance under well-specified and misspecified models. We found that the proposed statistics could generally isolate the model parameter violating measurement invariance, so long as the violating parameter is included in the model.

In the remainder of the paper, we first compare the use these tests to the use of traditional tests in practice. We then discuss test extension to other fit functions and to other specialized models.

6.1. APPLICATIONS

Many of the applications in this volume, along with many measurement invariance applications in general, focus on testing across unordered categories such as nations or gender. As discussed earlier in this paper, the score-based tests for unordered categories are equivalent to the usual likelihood ratio test. Given a measurement invariance violation across these unordered categories, however, researchers typically wish to know why the violation occurred. At this point, researchers may examine education level, socioeconomic status, income levels, and so on across the unordered categories. These variables are often ordinal or continuous in nature, so that the family of tests described in this paper are applicable. This is a first step toward describing why measurement invariance violations occur, as opposed to simply detecting measurement invariance violations. The tests described here are convenient for this purpose, as they do not require a new model to be estimated for each ordinal variable. Instead, each ordinal variable defines an ordering of observations, which in turn yields a test statistic that is specific to that ordinal variable.

6.2. EXTENSION

In this paper, we focused on testing for measurement invariance in factor analysis models that assume multivariate normality and that are estimated via maximum likelihood (ML). The family of tests described here generally apply to estimation methods that maximize/minimize a fit function, however (see Zeileis and Hornik, 2007), so they are potentially applicable to alternative SEM discrepancy functions such as generalized least squares (e.g., Browne and Arminger, 1995). Score calculation for these

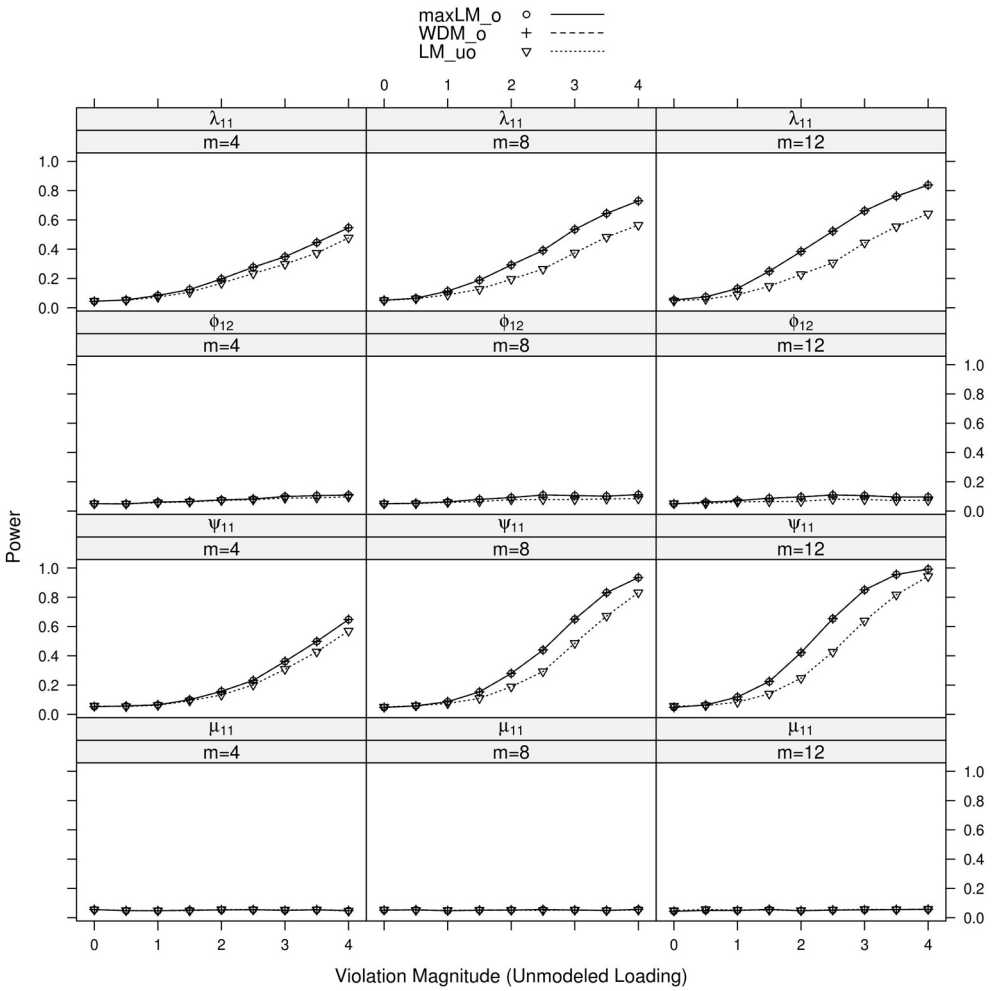


FIGURE 7 | Simulated power curves for max LM_o, WDM_o, and LM_{uo} across three levels of the ordinal variable *m* and measurement invariance violations of 0–4 standard errors (scaled by $\sqrt{n}$), Simulation 2. The parameter violating measurement invariance is the unmodeled loading. Panel labels denote the parameter(s) being tested and the number of levels of the ordinal variable *m*.

alternative discrepancy functions has not been implemented (to our knowledge), though the calculation could be implemented. Test statistic calculation and inference would then proceed in exactly the same manner as the calculation and inference illustrated in this paper. Study of the proposed tests' application to larger SEMs is warranted.

In addition to alternative fit functions, the tests can be extended to other models estimated via ML. Of primary relevance to the topic of measurement invariance, the tests can be extended to item response models to examine differential item functioning. In particular, Strobl et al. (2014) studied application of these tests to the Rasch model, using them as the basis of a recursive partitioning procedure that segments subgroups of individuals who exhibit DIF. Further study and extension of these tests for IRT are warranted.

COMPUTATIONAL DETAILS

All results were obtained using the R system for statistical computing (R Core Team, 2013), version 3.1.0, employing the add-on package *lavaan* 0.5–16 (Rosseel, 2012) for fitting of the factor analysis models and *strucchange* 1.5–0 (Zeileis et al., 2002; Zeileis, 2006) for evaluating the parameter instability tests. R and both packages are freely available under the General Public License 2 from the Comprehensive R Archive Network at http://CRAN.R-project.org/. R code for replication of our results is available at http://semtools.R-Forge.R-project.org/ and also in an online supplement to this article.

ACKNOWLEDGMENT

This work was supported by National Science Foundation grant SES-1061334.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpsyg.2014.00438/abstract

The accompanying zip file contains R code for replication of all analyses and simulations from the article. File descriptions appear below.

- mz-frontiers.R: Model estimation functions for simulations.
- sim-frontiers.R: Functions for data generation, power evaluation, and power summaries.
- replication-frontiers.R: Code for the tutorial and simulations, utilizing the other two files.

REFERENCES

Bauer, D. J., and Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychol. Methods* 14, 101–125. doi: 10.1037/a0017642

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238

Bentler, P. M., and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* 88, 588–606. doi: 10.1037/0033-2909.88.3.588

Browne, M. W., and Arminger, G. (1995). "Specification and estimation of mean- and covariance-structure models," in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, eds G. Arminger, C. C. Clogg, and M. E. Sobel (New York, NY: Plenum Press), 185–249. doi: 10.1007/978-1-4899-1292-3_4

Dolan, C. V., and van der Maas, H. L. J. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika* 63, 227–253. doi: 10.1007/BF02294853

Froh, J. J., Fan, J., Emmons, R. A., Bono, G., Huebner, E. S., and Watkins, P. (2011). Measuring gratitude in youth: assessing the psychometric properties of adult gratitude scales in children and adolescents. *Psychol. Assess.* 23, 311–324. doi: 10.1037/a0021590

Hansen, B. E. (1997). Approximate asymptotic *p* values for structural-change tests. *J. Bus. Econ. Stat.* 15, 60–67. doi: 10.2307/1392074

Hjort, N. L., and Koning, A. (2002). Tests for constancy of model parameters over time. *Nonparamet. Stat.* 14, 113–132. doi: 10.1080/10485250211394

Hothorn, T., and Zeileis, A. (2008). Generalized maximally selected statistics. *Biometrics* 64, 1263–1269. doi: 10.1111/j.1541-0420.2008.00995.x

Lubke, G. H., and Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychol. Methods* 10, 21–39. doi: 10.1037/1082-989X.10.1.21

McCullough, M. E., Emmons, R. A., and Tsang, J.-A. (2002). The grateful disposition: a conceptual and empirical topography. *J. Pers. Soc. Psychol.* 82, 112–127. doi: 10.1037/0022-3514.82.1.112

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Merkle, E. C., Fan, J., and Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*. doi: 10.1007/s11336-013-9376-7. [Epub ahead of print].

Merkle, E. C., and Zeileis, A. (2013). Tests of measurement invariance without subgroups: a generalization of classical methods. *Psychometrika* 78, 59–82. doi: 10.1007/s11336-012-9302-4

Molenaar, D., Dolan, C. V., Wicherts, J. M., and van der Mass, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence* 38, 611–624. doi: 10.1016/j.intell.2010.09.002

Muthén, B., and Asparouhov, T. (2013). *BSEM Measurement Invariance Analysis.* Technical Report, Mplus Web Note 17.

Purcell, S. (2002). Variance components models for gene-environment interaction in twin analysis. *Twin Res.* 5, 554–571. doi: 10.1375/twin.5.6.572

R Core Team (2013). *R: a Language and Environment for Statistical Computing*. R Foundation for Statistical Computing (Vienna, Austria). Available online at: http://www.R-project.org/

Rosseel, Y. (2012). *lavaan*: an R package for structural equation modeling. *J. Stat. Softw.* 48(2), 1–36. Available online at: http://www.jstatsoft.org/v48/i02/

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: a unified approach. *Psychometrika* 54, 131–151. doi: 10.1007/BF02294453

Strobl, C., Kopf, J., and Zeileis, A. (2014). A new method for detecting differential item functioning in the Rasch model. *Psychometrika*. doi: 10.1007/s11336-013-9388-3. [Epub ahead of print].

van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *Eur. J. Dev. Psychol.* 9, 486–492. doi: 10.1080/17405629.2012.686740

Zeileis, A. (2006). Implementing a class of structural change tests: an econometric computing approach. *Comput. Stat. Data Anal.* 50, 2987–3008. doi: 10.1016/j.csda.2005.07.001

Zeileis, A., and Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Stat. Neerland.* 61, 488–508. doi: 10.1111/j.1467-9574.2007.00371.x

Zeileis, A., Leisch, F., Hornik, K., and Kleiber, C. (2002). *strucchange*: an R package for testing structural change in linear regression models. *J. Stat. Softw.* 7(2), 1–38. Available online at: http://www.jstatsoft.org/v07/i02/

Zeileis, A., Strobl, C., and Wickelmaier, F. (2013). psychotools*: Infrastructure for Psychometric Modeling*. R package version 0.1-5. Available online at: http://CRAN.R-project.org/package=psychotools

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 04 November 2013; accepted: 25 April 2014; published online: 30 May 2014.
*Citation: Wang T, Merkle EC and Zeileis A (2014) Score-based tests of measurement invariance: use in practice. Front. Psychol. **5**:438. doi: 10.3389/fpsyg.2014.00438*
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology.

frontiers in Psychology

METHODS
published: 04 June 2015
doi: 10.3389/fpsyg.2015.00733

The comparability of the universalism value over time and across countries in the European Social Survey: exact vs. approximate measurement invariance

Florian Zercher[1]*, *Peter Schmidt*[1], *Jan Cieciuch*[2,3] and *Eldad Davidov*[4]

[1] Department of Political Science, University of Giessen, Giessen, Germany, [2] University Research Priority Program "Social Networks", University of Zürich, Zürich, Switzerland, [3] Institute of Psychology, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, Poland, [4] Institute of Sociology, University of Zürich, Zürich, Switzerland

OPEN ACCESS

Edited by:
Mike W.-L. Cheung,
National University of Singapore,
Singapore

Reviewed by:
Suzanne Jak,
Utrecht University, Netherlands
Rens Van De Schoot,
Utrecht University, Netherlands

***Correspondence:**
Florian Zercher,
Department of Political Science,
University of Giessen,
Karl-Glöcknerstrasse 21 E, 35394
Giessen, Germany
florian.zercher@sowi.uni-giessen.de

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: 11 October 2014
Accepted: 17 May 2015
Published: 04 June 2015

Citation:
Zercher F, Schmidt P, Cieciuch J and Davidov E (2015) The comparability of the universalism value over time and across countries in the European Social Survey: exact vs. approximate measurement invariance. Front. Psychol. 6:733. doi: 10.3389/fpsyg.2015.00733

Over the last decades, large international datasets such as the European Social Survey (ESS), the European Value Study (EVS) and the World Value Survey (WVS) have been collected to compare value means over multiple time points and across many countries. Yet analyzing comparative survey data requires the fulfillment of specific assumptions, i.e., that these values are comparable over time and across countries. Given the large number of groups that can be compared in repeated cross-national datasets, establishing measurement invariance has been, however, considered unrealistic. Indeed, studies which did assess it often failed to establish higher levels of invariance such as scalar invariance. In this paper we first introduce the newly developed approximate approach based on Bayesian structural equation modeling (BSEM) to assess cross-group invariance over countries and time points and contrast the findings with the results from the traditional exact measurement invariance test. BSEM examines whether measurement parameters are approximately (rather than exactly) invariant. We apply BSEM to a subset of items measuring the universalism value from the Portrait Values Questionnaire (PVQ) in the ESS. The invariance of this value is tested simultaneously across 15 ESS countries over six ESS rounds with 173,071 respondents and 90 groups in total. Whereas, the use of the traditional approach only legitimates the comparison of latent means of 37 groups, the Bayesian procedure allows the latent mean comparison of 73 groups. Thus, our empirical application demonstrates for the first time the BSEM test procedure on a particularly large set of groups.

Keywords: European Social Survey, approximate vs. exact measurement invariance, Portrait Value Questionnaire, universalism, Bayesian estimation, cross-national research, repeated cross-sections

Over the last decades, considerable research on values has taken place (Hitlin and Piliavin, 2004). These theoretical and empirical research contributions have been inspired especially by Inglehart and his colleagues (Inglehart, 1977; Inglehart and Welzel, 2005) and Schwartz and colleagues (Schwartz, 2003; Schwartz et al., 2012). Inglehart's value measurement instruments have been implemented in the World Value Survey (WVS), whereas a short version of Schwartz's Portrait Values Questionnaire (PVQ) with 21 items has been included in the European Social Survey (ESS).

Comparisons of the two theoretical conceptions and the measurement instruments based on them were undertaken by Datler et al. (2013) and Beckers et al. (2012).

To date, the PVQ has been the object of extensive comparative research in the social sciences. Studies have focused, for example, on the relation between values and political behavior, left-right orientation, attitudes toward immigration, attitudes toward homosexuality and sociodemographic characteristics (Davidov et al., 2008, 2014b; Piurko et al., 2011; Meuleman et al., 2012; Schwartz et al., 2012; Kuntz et al., 2015) by making use of increasingly available cross-national data sources, such as the ESS or the WVS. The cross-national orientation in the study of values offered the advantage of introducing a stricter test of propositions (Popper, 2005), thereby expanding our knowledge about the validity of theories in different societies and allowing us to acquire insights into macro-micro effects (Opp, 2011). However, in comparative research, the issue of comparability across countries must be addressed (Davidov et al., 2014a). Respondents in different countries may understand survey questions in various ways (Latcheva, 2011; Braun et al., 2013) or respond in systematically different ways to the same questions (Harkness et al., 2010). This may lead to biased means, factor loadings and regression coefficients. Therefore, the assumption of cross-cultural measurement invariance needs to be tested (Meredith, 1993; Vandenberg and Lance, 2000; Davidov and Siegers, 2010; Millsap, 2011; Sarrasin et al., 2012; van de Schoot et al., 2012; Davidov et al., 2014a).

Davidov et al. (2008) and Davidov (2008, 2010) tested the measurement invariance properties of values across countries in three rounds of the ESS and could establish only metric invariance within the rounds across most countries and longitudinal scalar invariance within countries[1]. However, it remains to be answered if value measurements are invariant both across countries and over time and whether such an extensive test is feasible with real data. After all, various researchers who use values as explanatory or as explained constructs wish to test propositions referring simultaneously to different countries ("the cross-cultural aspect") and time points ("the dynamic aspect"). Such an endeavor requires that measurement invariance is given simultaneously over time and across countries. However, such a measurement invariance test has not been performed in the past. Moreover, such a test becomes increasingly important considering the continuous growth in the number of countries *and* time points in the large data-generating programs mentioned before. Thus, our research question is whether it is feasible to test and establish measurement invariance across a very large number of groups.

In the current study we would like to focus on the universalism value because it is the only value which was measured in the PVQ-21 with three (rather than only two) items, thus allowing us to control for all forms of random and nonrandom measurement errors (Bollen, 1989). Furthermore, this universalism scale has also been used in a considerable number of empirical studies using ESS data (Jowell et al., 2007; Beierlein et al., 2012; Davidov et al., 2012; Saris et al., 2013) and other datasets (Schwartz et al., 2012; van de Schoot et al., 2012). We will examine its simultaneous comparability across 15 countries and six time points using the new procedure for assessing approximate invariance using Bayesian estimation (van de Schoot et al., 2013). To the best of our knowledge, no previous study has assessed invariance across so many groups simultaneously[2]. We will demonstrate the application of the two approaches on the same large set of time/country groups. Given previous findings, we expect to find metric invariance at best for the universalism scale but no scalar invariance across countries using the traditional exact method. However, we expect to establish scalar invariance at least for a subset of countries using the approximate approach.

We begin by briefly presenting the traditional exact approach and then describe the new approximate approach to test for measurement invariance across groups. Next, we describe our data and the three items that measure universalism. In the empirical part we report the results of the two approaches to test for invariance. We finalize with a discussion of the pros and cons of the traditional exact approach vs. the approximate approach to test for measurement invariance in cross-national research.

The Traditional Approach to Measurement Invariance Testing: Multi-Group Confirmatory Factor Analysis (MGCFA)

Multi-group confirmatory factor analysis (Jöreskog, 1971; Bollen, 1989; Brown, 2006) has been the most common method used to test for measurement invariance. There are three distinct and hierarchically ordered levels of measurement invariance. Each level is defined by the parameters constrained to be equal across groups. The first and lowest level is configural invariance (Horn and McArdle, 1992; Meredith, 1993; Vandenberg and Lance, 2000). Configural invariance requires that each construct is measured by the same items. The second level is metric invariance, and it guarantees that the measured construct essentially has the same meaning in the different groups under study. Full metric invariance is tested by constraining the factor loadings to be equal across the groups to be compared (Vandenberg and Lance, 2000). If full metric invariance is established, a one-unit increase in the latent construct has the same meaning across groups. Subsequently, covariances and unstandardized regression coefficients may be meaningfully compared across samples (Steenkamp and Baumgartner, 1998). However, it is still uncertain whether the construct is measured on the same scale (Horn and McArdle, 1992; Steenkamp and Baumgartner, 1998; Vandenberg and Lance, 2000). Scalar invariance requires, in addition, that the intercepts are equal across groups. It is tested by constraining both the factor loadings and the intercepts to be equal across the groups to be compared (Vandenberg and Lance, 2000). If full scalar invariance

[1]For an invariance test of a new scale to measure human values, see Cieciuch et al. (2014a,b).

[2]In the study of van de Schoot et al. (2013), only a small number of groups was studied. Cieciuch et al.'s (2014a) studies contained eight groups, and Davidov et al. (2015) contained 15 groups in six separate tests. None of these studies performed a simultaneous test over countries and time points, which would have led to a much higher number of groups.

is established, also the means may be meaningfully compared across groups (Steenkamp and Baumgartner, 1998).

Below, the corresponding three sets of constraints for the three levels of invariance are defined for a particular item in a one-factor case for individual i in group j (see Muthén and Asparouhov, 2013).

$$\begin{aligned} \text{Configural invariance } y_{ij} &= \nu_j + \lambda_j f_{ij} + \varepsilon_{ij} \\ E(f_i) &= \alpha_j,\ V(f_j) = \psi_j \end{aligned} \quad (1)$$

Where v is a measurement intercept, λ is a factor loading, f is a factor with mean α and variance Ψ, and ε is a residual with mean zero and variance θ, uncorrelated with f. The configural model has subscript j for both intercepts and loadings.

$$\begin{aligned} \text{Metric invariance } y_{ij} &= \nu_j + \lambda f_{ij} + \varepsilon_{ij} \\ E(f_i) &= \alpha_j = 0,\ V(f_j) = \psi_j \end{aligned} \quad (2)$$

The metric model drops the subscript j for the loadings because they are assumed to be equal.

$$\begin{aligned} \text{Scalar invariance } y_{ij} &= \nu + \lambda f_{ij} + \varepsilon_{ij} \\ E(f_i) &= \alpha_j,\ V(f_j) = \psi_j \end{aligned} \quad (3)$$

The scalar model drops the subscript j for both intercepts and loadings because they are assumed to be equal[3].

In practice, it is particularly difficult to reach full scalar invariance. Variations in the way respondents react to questions or systematic response biases such as social desirability or acquiescence (Billiet et al., 2003; Oberski et al., 2012), which may be individually or culturally determined, could possibly distort responses to the extent that scalar invariance will not exist in most empirical applications (Davidov et al., 2014a). There have been basically two major approaches to handling the issue of measurement noninvariance (Jouha and Moustaki, 2013; van de Schoot et al., 2013; Davidov et al., 2014a):

(1) Ignoring it. This is what the overwhelming majority of researchers have done as is evident in publications using cross-national and multigroup data, repeated cross-sections and panel data (see Davidov et al., 2014a). This line of literature has typically used sum scores instead of first testing whether the assumption of invariance can be supported by the data. As Steinmetz (2013) demonstrated in a Monte Carlo study, the use of sum scores is not an adequate procedure without invariance testing, as sum score differences are only warranted in conditions of full measurement invariance.

(2) Byrne et al. (1989) and Steenkamp and Baumgartner (1998) proposed the concept of *partial invariance* as a sufficient condition for meaningful cross-group comparisons. This approach has become a standard approach among various researchers. Partial invariance is given if the parameters of *at least* two indicators per construct (i.e., loadings for partial metric invariance and loadings plus intercepts for partial scalar invariance) are equal across groups. Several scholars rely on partial invariance when comparing countries, cultures or other units of analysis. However, even partial scalar invariance may often be rejected.

Three common procedures in the MGCFA literature which rely on global fit measures have been proposed to evaluate whether measurement invariance is established:

(1) To rely on the chi-square difference test and compare the configural, metric and scalar invariance models, which form nested models (Jöreskog, 1978; Bollen, 1989; Meredith, 1993; Brown, 2006). According to this procedure, the chi-square difference test is used to assess the correctness of the model. However, the use of the chi-square difference test has been criticized because of its sensitivity to sample size (among other reasons) (Jöreskog, 1993; Cheung and Rensvold, 2002).

(2) To use cut-off values for the *difference* in the comparative fit index (CFI), the root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMR) (Chen, 2007; for alternative cut-off values see Meade et al., 2008). According to this procedure, if the change in model fit is smaller than the criteria proposed in the literature, measurement invariance for that level is established. According to the results of Chen's (2007) simulation study, the following recommendations have been proposed:

   (a) If the sample size is larger than 300, metric noninvariance is indicated by a change in CFI larger than 0.01 supplemented by a change in the RMSEA larger than 0.015 or a change in SRMR larger than 0.03 compared with the configural invariance model.

   (b) Scalar noninvariance is evidenced by a change in CFI larger than 0.01 supplemented by a change in RMSEA larger than 0.015 or a change in SRMR larger than 0.01 compared with the metric invariance model.

(3) The third procedure suggests employing the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) information theoretic measures to compare the configural, metric and scalar invariance models (Kass and Raftery, 1995). Following the criteria proposed by Kass and Raftery (1995), a very strong difference is indicated when the AIC or BIC difference is greater than 1[4].

Since empirical tests often fail to establish measurement invariance based on these criteria, it has been argued that the criteria for testing measurement invariance may be too strict (Muthén and Asparouhov, 2013) and that more liberal criteria should be used to assess approximate (rather than exact) measurement invariance.

[3] In the Analytical Strategy section we shortly describe our approach to identify the models.

[4] A more detailed analysis of the issue of robustness against violations of metric and scalar invariance is given in Jouha and Moustaki (2013), Oberski (2014), and Meuleman (2012). See also Saris et al. (2009) for an alternative procedure to assess whether exact measurement invariance is given which relies on identifying local misspecifications while taking the power of the test into account. Furthermore, Thompson and Green (2013) argue that it might be better to rely on theory and past empirical findings and to be less dependent on empirical methods like the global fit measures and the modification indices when deciding whether to accept or reject a given level of invariance. This issue has not been settled yet.

The Bayesian Approach to Test for Approximate Measurement Invariance

Recently, Muthén and Asparouhov (2013) and van de Schoot et al. (2013) proposed an alternative approach to test for measurement invariance by applying approximate Bayesian measurement invariance testing. The exact procedure, which constrains factor loadings and intercepts to be exactly equal to establish measurement invariance, is very restrictive and rarely establishes invariance (Jouha and Moustaki, 2013; van de Schoot et al., 2013). Approximate measurement invariance permits "small" differences between parameters (van de Schoot et al., 2013). The parameters specified in a Bayesian approach are considered to be variables, and their distribution is described by priors. The assignment of prior distributions to unknown parameters reflects the researcher's uncertainty about them regardless of whether one conceives of a parameter as having one true value or not (Levy and Choi, 2013). Such uncertainty may be applied for various parameters both in single-group CFA and MGCFA. In invariance testing one may assume that differences between parameters (factor loadings, intercepts) are approximately equal. Thus, we would allow the introduction of some uncertainty by specifying a small variance of, for example, 0.01 or 0.05 around the difference in factor loadings or intercepts (van de Schoot et al., 2013).

Figure 1 delineates the difference between the traditional exact approach to test for measurement invariance and the Bayesian approximate approach. In the traditional exact approach, the differences of factor loadings (λ) or intercepts (ν) between groups are assumed to be exactly zero, while in the Bayesian approach the differences are assumed to be approximately zero with a mean of zero and some small variance delta (δ). Thus, we allow small variations in a given interval between the parameters as part of the measurement model[5] (see also Kruschke et al., 2012; Muthén and Asparouhov, 2012, 2013; Levy and Choi, 2013). Simulations suggest that "small" variations may be allowed without risking invalid conclusions in comparative research (van de Schoot et al., 2013).

The difference between the traditional exact approach and the Bayesian approximate approach is also evident in the definitions of the confidence interval (used in the traditional exact approach) and the credibility interval (CI) (used in the Bayesian approximate approach). The confidence interval over an infinite number of samples taken from the population expresses that 95% of these contain the true population value. By way of contrast, the CI expresses that there is a 95% probability that the population value is within the limits of the interval.

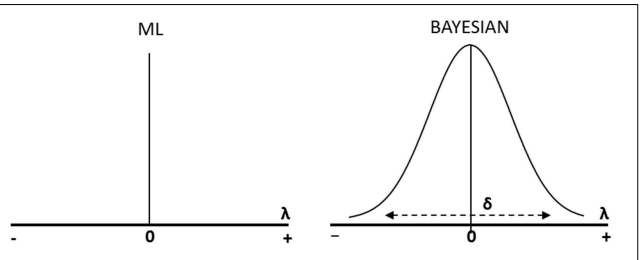


FIGURE 1 | Difference in parameter estimation between Maximum Likelihood (ML) and the Bayesian approach (see Muthén and Asparouhov, 2013; van de Schoot et al., 2013).

A number of fit measures have been proposed to specifically assess Bayesian models (Gelman, 2003, 2013; Levy, 2011). These fit measures can detect if the actual deviations are larger than those allowed by the researcher in the prior distribution. First, the model fit can be evaluated based on the posterior predictive probability value (ppp). The ppp is computed by comparing two types of information: the discrepancy between the model and the observed data and the discrepancy between the model and the posterior predicted data (Levy and Choi, 2013, p. 597)[6]. According to Muthén and Asparouhov (2012) and van de Schoot et al. (2013), the ppp value of a model that fits the data should be nonsignificant, and if it is around 0.50, it indicates a well-fitting model.

A second fit measure refers to the CI for the difference between the observed and the replicated chi-square values. According to Muthén and Asparouhov (2012) and van de Schoot et al. (2013), the CI should contain zero. Finally, the BIC (Schwarz, 1978) and the deviance information criterion (DIC) (Spiegelhalter et al., 2002) were also proposed for the assessment of model comparison in a Bayesian framework (Kass and Raftery, 1995). BIC is computed using the following formula:

$$BIC = -2\ell\left(\hat{\theta}|X\right) + p * \ln(n) \tag{4}$$

where $\ell\left(\hat{\theta}|X\right)$ is the maximized log-likelihood, p is the number of parameters, and n is the number of observations. Building on this tradition of comparing values of information criteria, Spiegelhalter et al. (2002) introduced the DIC:

$$DIC = \overline{D(\theta)} + p_D = 2\overline{D(\theta)} - D(\overline{\theta}) + 2p_D \tag{5}$$

where $D(\theta)$ is the posterior mean of the deviation (negative of twice the log-likelihood function), p_D is a complexity measure defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean, $D(\overline{\theta})$[7].

Testing for approximate measurement invariance consists of two steps. The first identifies the noninvariant parameters while fitting the model to data. Noninvariant parameters are those parameters which are found to be outside of the "wiggle room" allowed for the parameter differences. In the second step the noninvariant parameters are freed and the model is recalculated (Muthén and Asparouhov, 2013; van de Schoot et al., 2013). In the next section we are going to provide a practical application by demonstrating a test for approximate invariance using ESS data.

[5]Whether and to what extent our analysis procedure corresponds with the common concept of using prior knowledge in the same way as in Bayesian statistics is debatable, since our priors actually correspond to an assumption testing of approximate invariance rather than strictly to prior knowledge.

[6]This procedure corresponds to the comparison between the observed variance-covariance matrix (S) and the expected variance-covariance matrix ($\sum$) using maximum likelihood estimation in structural equation modeling (Bollen, 1989).

[7]For a discussion of other fit measures for Bayesian SEM, see Kaplan (2014) and Levy and Choi (2013). Indeed, as Levy and Choi(2013, p. 599) argue, little research has been conducted on the relative merits and limitations of these fit measures to evaluate model comparisons in BSEM.

TABLE 1 | ESS sample sizes for the selected 15 countries over six ESS rounds (2002–2012).

	1st Round (2002/3)	2nd Round (2004/5)	3rd Round (2006/7)	4th Round (2008/9)	5th Round (2010/11)	6th Round (2012/13)	*N*
Belgium	1899	1778	1798	1760	1704	1869	10,808
Switzerland	2040	2141	1804	1819	1506	1493	10,803
Germany	2919	2870	2916	2751	3031	2958	17,445
Denmark	1506	1487	1505	1610	1576	1650	9334
Spain	1729	1663	1876	2576	1885	1889	11,618
Finland	2000	2022	1896	2195	1878	2197	12,188
United Kingdom	2052	1897	2394	2352	2422	2286	13,403
Hungary	1685	1498	1518	1544	1561	2014	9820
Ireland	2046	2286	1800	1764	2576	2628	13,100
Netherlands	2364	1881	1889	1778	1829	1845	11,586
Norway	2036	1760	1750	1549	1548	1624	10,267
Poland	2110	1716	1721	1619	1751	1898	10,815
Portugal	1511	2052	2222	2367	2150	2151	12,453
Sweden	1999	1948	1927	1830	1497	1847	11,048
Slovenia	1519	1442	1476	1286	1403	1257	8383
N	29,415	28,441	28,492	28,800	28,317	29,606	173,071

Method and Data

For the analysis we employ data from the ESS measuring the universalism value (Schwartz, 2003; Schwartz et al., 2012)[8]. The ESS is a biannual cross-national European survey that is administered to representative samples from approximately 30 countries. Since its inception in 2002/2003, it has included questions that measure values in its core module. These questions have been repeated in each round and used extensively in cross-national research. In the present analysis we have included 15 countries which participated in all six rounds. **Table 1** presents the sample sizes for each country/time point combination between 2002 and 2012.

Three items were used to measure the universalism value. Respondents were presented with a descriptive portrait of a person (gender matched), and they were requested to indicate to what extent they were similar to this person. The response scale ranged from 1 (*very much like me*) to 6 (*not like me at all*). These responses were reversed so that higher scores represented greater similarity to enable a more straightforward interpretation of the scores. The correlations between items were considerable and ranged approximately between 0.3 and 0.4. The rate of missing values for these items ranged from 4.0 to 4.2% only for each country/time point combination. **Table 2** presents the item formulations.

TABLE 2 | Formulation of universalism items.

"Now I will briefly describe some people. Please listen to each description and tell me how much each person is or is not like you. Use this card for your answer…"
Universalism Item1–"…She/he thinks it is important that every person in the world should be treated equally. She/he believes everyone should have equal opportunities in life."
Universalism Item2–"…It is important to her/him to listen to people who are different from her/him. Even when she/he disagrees with them, she/he still wants to understand them."
Universalism Item3–"…She/he strongly believes that people should care for nature. Looking after the environment is important to her/him."

Analytical Strategy

Testing for Exact (Full or Partial) Invariance

In the first step we performed six MGCFAs (one for each round) across 15 countries, and after that, the analysis was performed on all 15 countries and six rounds (with a total of 90 groups) simultaneously. In both cases, the full information maximum likelihood (FIML) procedure was used to deal efficiently with the problem of missing values (Schafer and Graham, 2002). We used the robustified maximum likelihood estimation procedure to deal with the ordered categorical character of the data[9].

Each analysis contained assessments for configural, metric and scalar invariance, with the corresponding constraints for each level of the measurement invariance[10]. In a second step, when full measurement invariance was not established, we tried to assess partial measurement invariance. In order to establish partial scalar invariance (where at least two items are constrained to be exactly equal), the intercept of only one item was released, because partial scalar invariance requires that parameters of at least two items are constrained to be equal across all groups.

[8]The raw data is available at the official site of the European Social Survey: http://www.europeansocialsurvey.org/downloadwizard.

[9]Only standard errors and chi-square differ between MLR and FIML.

[10]To identify the model we used the marker variable method (MVM; see Little et al., 2006). We constrained the factor loading of one item to one and its intercept to zero. To test the robustness of our findings, we re-ran the model two more times, and each time with a different item as the marker item. The results remained essentially the same.

Testing for Approximate Invariance

Following Muthén and Asparouhov (2013) and van de Schoot et al. (2013), we ran models with informative priors with a mean of zero and variances of 0.005, 0.01, 0.05, and 0.5 for the differences between factor loadings or intercepts across groups[11]. Next, we identified in each model with the different priors those factor loadings and intercepts which were different. In the next step we freed all parameters which were considerably different across groups and left the informative priors for all the other equality parameters intact (Muthén and Asparouhov, 2013). **Table 3** summarizes the steps undertaken in each approach. These analyses were conducted on all ESS rounds and countries simultaneously.

Results

The Traditional Exact Approach

Table 4 presents the global fit measures of the accepted models after dropping countries using the traditional exact approach. The first part of the table presents the global fit measures of the accepted model in each round separately. The last part of the table presents the global fit measures for the accepted model in the simultaneous analysis across countries and rounds. After releasing the equality constraint on the intercept that had the highest modification index in most country/time point combinations (Byrne et al., 1989), we identified in the simultaneous analysis 53 country/time point combinations in which at least two items were noninvariant. These country/time point combinations had to be dropped from further analysis because, for these units, even partial invariance could not be established. For example, the items which measured the importance to understand different people and to take care of the environment were scalar noninvariant in Switzerland and Denmark at all measurement time points. Consequently, we dropped these countries from further analysis. Thus, in total, 37 of the country/time point combinations displayed partial exact scalar invariance[13].

Furthermore, we employed AIC and BIC comparisons of the metric invariance and partial scalar invariance models (see **Table 5**) in the separate analyses for each round and in the simultaneous analysis. Following the criteria proposed by Kass and Raftery (1995) to compare BIC differences, we can conclude that all differences between the metric and the partial scalar model, in a reduced number of countries, are very large.

The results have two important implications. On the one hand, findings of partial scalar invariance allow meaningful mean comparison across 37 country/time point combinations for the universalism construct. However, it is discouraging to find out that mean comparisons of the universalism value may be problematic in so many of the country/time point combinations. Next, we turn to the approximate invariance test.

TABLE 3 | Analytical steps for the exact and the approximate measurement invariance approaches.

	Traditional exact approach	Approximate approach
Steps	1. Configural model 2. Metric model 3. Scalar model 4. Partial scalar model	1. Setting different informative priors for the cross-group differences of loadings and intercepts 2. Releasing (approximate) equality constraints (of loadings and intercepts) that are not supported by the data
Additional steps[12]	5. Deleting groups which are not fully or partially scalar invariant	3. Deleting groups which are not fully or partially approximately invariant

As metric invariance could be established in the exact approach, we did not need to fall back to partial metric invariance.

TABLE 4 | Global fit measures of the traditional exact approach.

	Chi^2(df)	RMSEA	SRMR	CFI	Countries/ Timepoints[14]
ROUND 1					
Partial scalar	64.89 (24)	0.029	0.029	0.985	8
ROUND 2					
Partial scalar	53.28 (28)	0.022	0.027	0.992	9
ROUND 3					
Partial scalar	53.78 (27)	0.024	0.033	0.988	8
ROUND 4					
Partial scalar	87.43 (24)	0.040	0.041	0.978	8
ROUND 5					
Partial scalar	90.10 (21)	0.044	0.039	0.972	7
ROUND 6					
Partial scalar	69.26 (21)	0.034	0.036	0.980	7
COUNTRIES AND ROUNDS SIMULTANEOUSLY					
Partial scalar	348.23 (126)	0.031	0.035	0.983	37[15]

*RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; CFI, comparative fit index; the partial scalar model corresponds to step 5 in **Table 3**.*

[11] When running the Bayesian procedure, we first ran a model where the difference between factor loadings or intercepts across groups has a normal distribution prior with a mean of 0 and a very large variance of 10^{10} (the so-called noninformative prior). This allows us to firstly detect whether there are any calculation problems in the Bayesian analysis (van de Schoot et al., 2013).

[12] After we were unable to achieve partial measurement invariance using the common ways of model fitting, we had to delete countries/time points (groups) based on the modification indices for the exact approach and based on the single group ppp for the approximate approach.

[13] Discussing possible explanations why specific countries are not comparable to others is beyond the scope of the present study. See Davidov et al. (2012) for using multilevel structural equation modeling for explaining noninvariance.

[14] For the single rounds this refers to countries; for all rounds this is combination of countries and time points.

[15] Countries still included are: Belgium 2002–2012; Spain 2002–2006; Finland 2006–2010; United Kingdom 2012; Hungary 2002–2008; Ireland 2008, 2010; Netherlands 2002–2012; Norway 2004–2012; Poland 2006; Portugal 2004–2008; Sweden 2012; Slovenia 2002, 2006.

TABLE 5 | AIC and BIC fit measures of the traditional exact approach[16].

		AIC	BIC
Round 1	Metric	232453.884	233335.682
	Partial scalar	133004.879	133373.601
Round 2	Metric	218452.710	219328.143
	Partial scalar	134813.330	135221.803
Round 3	Metric	222284.379	223163.765
	Partial scalar	106349.111	106687.021
Round 4	Metric	225469.593	226350.568
	Partial scalar	109976.943	110337.466
Round 5	Metric	226639.903	227520.419
	Partial scalar	98034.755	98344.903
Round 6	Metric	237036.130	237923.153
	Partial scalar	113273.097	113589.931
All rounds	Metric	1362329.608	1368665.132
	Partial scalar	537676.482	539559.803

TABLE 6 | Global fit measures for the approximate invariance test (mean = 0 and variance = 0.05).

	ppp	ppp after releasing misspecified parameters	CI after releasing misspecified parameters
90 groups	0.000	0.000	125.830–346.761
73 groups[19]	0.026	0.052	−10.834–171.115

ppp, posterior predictive probability; CI, credibility interval.

TABLE 7 | Correlations between latent means computed using sum scores (1), the exact (2) and the approximate (3) measurement invariance models for 73 county/time points.

	Sum scores (1)	Exact test[20] (2)	Approximate Bayesian test (3)
1	1		
2	0.997**	1	
3	0.851**	0.844**	1

** $p < 0.01$ *(pairwise deletion).*

The Bayesian Approximate Approach

Here, too, we first tested each round separately and then all rounds simultaneously[17]. Approximate measurement invariance across all countries was established in only two rounds (2002 and 2004). Next, as recommended by van de Schoot et al. (2013), we ran the model that included all time points and countries, using several prior variances to compare them. We released equality constraints on those loadings and intercepts which were different[18] . Finally, we deleted groups which were not approximately invariant. **Table 6** reports the results for the model with a prior of 0.05 (Muthén and Asparouhov, 2013; van de Schoot et al., 2013).

Accordingly, 73 countries/time points remained in the model. Thus, the results suggest that the exact and approximate measurement invariance approaches produce quite different findings. Whereas, partial approximate scalar measurement invariance was established in 73 ESS country/time point combinations, exact scalar measurement invariance was only established in 37 country/time point combinations. In other words, the approximate test allows us to perform mean comparisons of universalism across a very large set of countries and time points.

Mean Comparison

We compared the country means obtained from the MGCFA and Bayesian analyses with each other as well as with those based on the raw sum scores for the 73 comparable country/time point combinations. This was done by estimating mean scores based on the exact and approximate approaches and comparing them to each other and to those computed using the raw data. Finally, we estimated the correlation between the means computed in the country/time point combinations based on each of the three procedures.

As **Table 7** demonstrates, the correlation is highest between sum scores and the exact test (0.997), and the correlation between the Bayesian approximate test and the exact test (0.844) is lowest. Since the latent means from the approximate test are the only ones which rely on an acceptable model fit, we conclude that latent means based on the other approaches (the exact and the sum scores) are biased. **Figure 2** presents the differences in the means between the sum scores and the scores from the approximate approach on a scatter plot. If the scores in the two methods were equal, they would all be on the diagonal. Stated another way, increased distance from the diagonal indicates increased differences between the scores.

Conclusions may also be biased when sum scores are compared for the same country longitudinally. **Figure 3** presents the mean over time and within countries. For example, as **Figure 3** demonstrates, when comparing the sum scores in Poland, one would assume that the means considerably increased between 2002 and 2012. However, based on the approximate approach, the data show that there was no mean difference between 2002 and 2012 for the universalism value scores in Poland. By way of contrast, the sum scores indicate no mean difference between 2002 and 2012 in Ireland. However, according to the approximate test, there was a slight increase in the universalism mean in Ireland between the two rounds. We thus conclude that if a researcher would draw conclusions based on the composite scores, either to compare countries with each other

[16]The partial scalar model corresponds to step 5 in **Table 3**.

[17]An example of the syntax can be found in the Supplementary Material. We would like to thank Bengt and Linda Muthén very much for making it possible to run such a model in the Mplus 7.3 version (Muthén and Muthén, 1998–2014). Previous versions did not allow the inclusion of this number of groups.

[18]A detailed report of the results is beyond the scope of the present study and may be provided by the first author upon request.

[19]Countries/time points not included are Denmark 2002, 2004, 2010, 2012; Spain 2008, 2010, 2012; Finland 2002, 2004; United Kingdom 2010; Hungary 2008; Ireland 2012; Norway 2008; Poland 2008, 2010; Sweden 2012; Slovenia 2010.

[20]To illustrate the comparison, these latent means are based on the model with all countries from the exact test that did not achieve scalar invariance.

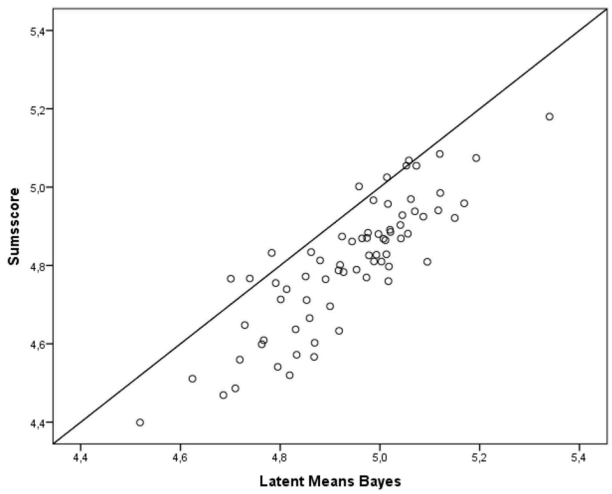


FIGURE 2 | Relationship between sum scores and scores based on the Bayesian estimation in 73 country/time point combinations.

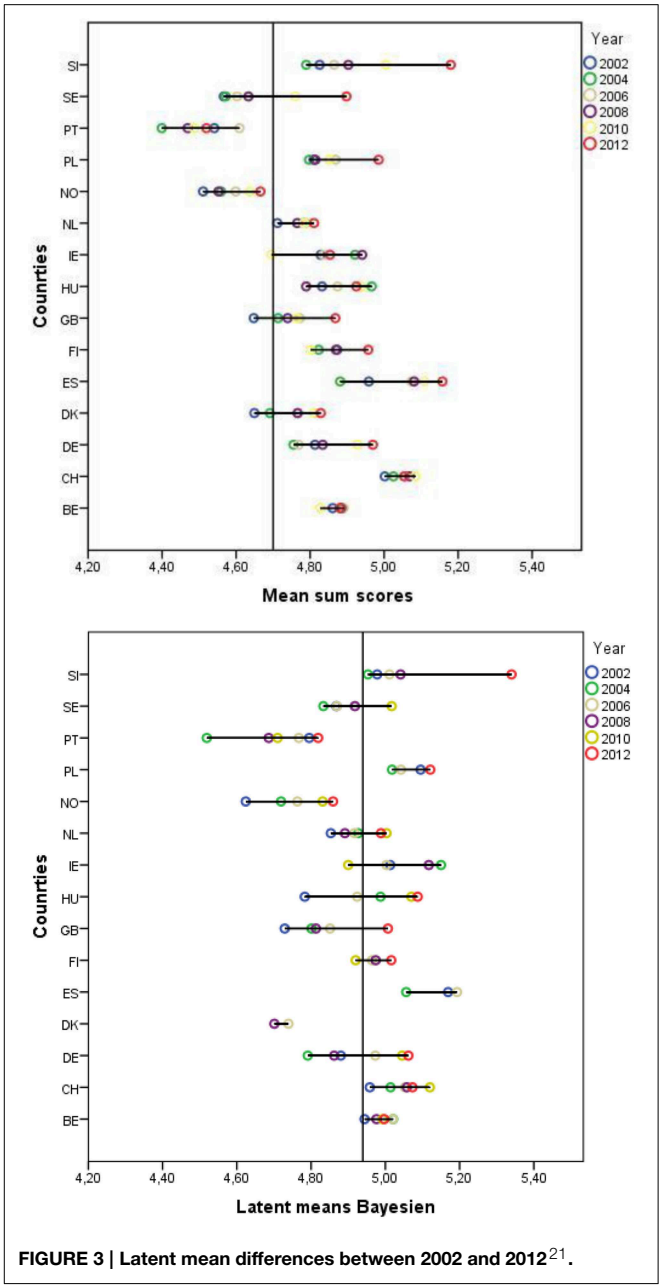


FIGURE 3 | Latent mean differences between 2002 and 2012[21].

or to compare scores within the same country and over time, they might be misled by the scores and reach wrong conclusions. In **Figure 3** one can see the variance of the latent means over the six time points. The length of the line shows the variation and the colored circles show the latent mean of universalism at in each round.

Finally, **Figure 4** displays the mean development of universalism over time in each of 15 countries and how this compares to the overall mean level of universalism across all countries and rounds.

To visualize the latent means over different time points and different countries, we split the countries into three groups comprising five countries each. The straight, dotted horizontal line is the mean over all country/time point groups. The graphs depicted in **Figure 4** suggest that the mean of universalism increases over time in most countries, while it remains more or less stable in Portugal, Ireland, Finland, and Belgium.

Summary and Conclusions

In most published cross-national studies, metric and scalar measurement invariance is implicitly assumed without testing this assumption. This may lead to biased mean comparisons and biased comparisons of covariances and regression coefficients (Vandenberg and Lance, 2000; Jouha and Moustaki, 2013; Oberski, 2014). However, the traditional estimation procedures used in MGCFA to test for measurement invariance and the corresponding global fit measures, especially in the case of scalar invariance assessments, mostly lead to a rejection of the assumption of even partial invariance. This often results in a considerable reduction in the number of countries and/or time points whose means can be meaningfully compared.

In the current study we assessed the comparability of the universalism value in six rounds of the ESS between 2002 and 2012 across all ESS countries, with 90 country/time point combinations in total. To the best of our knowledge, this is the first time in which so many groups are included in such a test. Using the traditional exact measurement invariance test procedure, metric invariance could be established across all country/time point combinations although partial scalar invariance could not, and we were required to drop almost two thirds of the countries/time points based on the reason that their mean scores on the scale might not be comparable.

[21] In the figure with the Bayesian latent means not all countries and time points are included. Countries/time points which are not included are Denmark 2000, 2002, 2010, 2012; Spain 2008, 2010, 2012; Finland 2000, 2002; United Kingdom 2010; Hungary 2008; Ireland 2012; Norway 2008; Poland 2008, 2010; Sweden 2012; Slovenia 2010.

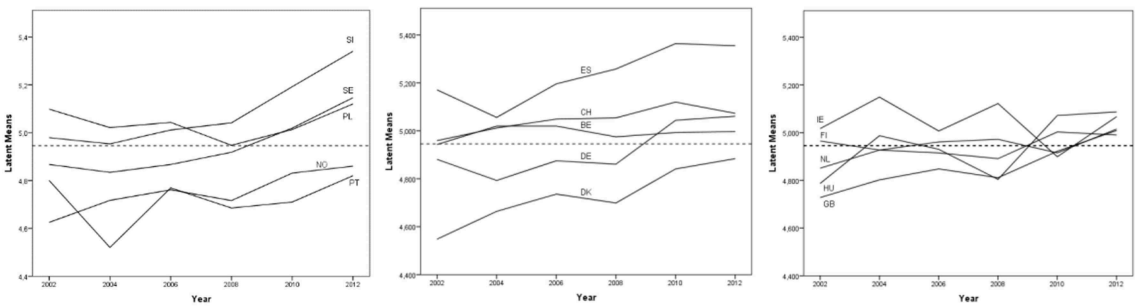


FIGURE 4 | Latent means over different time points and different countries[22].

To solve this problem we applied the newly proposed approximate measurement invariance procedure. In these analyses only 17 country/time point combinations had to be excluded. We could demonstrate that the assumption of (approximate) scalar invariance was tenable using this alternative procedure on the remaining countries. As a consequence, the latent means of universalism could be legitimately compared across many more countries and time points.

Having said that, we believe that the traditional exact approach should always be applied as a first step in invariance testing. After all, it could well be the case that measurements are exactly invariant and it is not necessary to apply approximate (rather than exact) constraints. Using only the exact approach may circumvent not only using the (technically more challenging) approximate approach but a practical problem we encountered while analyzing the data applying the approximate approach as well: Using it for so many groups with large sample sizes led to a computation time of between 12 and 16 h! However, where even partial *exact* measurement invariance does not hold, it would be useful to apply the approximate approach using Bayesian estimation (van de Schoot et al., 2013). This may be a relevant assessment especially in the case of comparisons of many groups such as in cross-national research with repeated cross-sections. As previous studies have demonstrated, in such cases it may be particularly difficult to establish full or partial (exact) scalar invariance.

It should be noted, however, that such a result in which so many country/time point combinations demonstrate approximate invariance may not necessarily be replicated with other data and other scales. Indeed, it could well be the case that both exact and approximate approaches fail to demonstrate cross-country and over time invariance. In other words, the approximate approach does not establish invariance where it is not given. It is, however, more liberal than the exact approach and may establish approximate invariance although the exact test fails to do so.

Future research may analyze various cross-national datasets with large samples to evaluate the approximate comparability of various scales and the practical usefulness of the approximate approach used here. In addition, it would be desirable if further simulation studies would be performed to evaluate which priors may be used in approximate invariance tests and which ppp values should be considered supportive for the assessed models. Such simulations could also explore how increasing the number of groups and the number of respondents in the groups may influence the results. This issue is particularly relevant because the number of groups (such as countries, cultural groups, language groups, etc.) in large data-generating programs such as the ESS, EVS, Eurobarometer, WVS, or the PISA study is continuously increasing. Furthermore, given that very often invariance cannot be established, it would be desirable if future studies would seek explanations for the absence of measurement invariance (see, e.g., Davidov et al., 2012, 2015). Finally, future research which includes a large number of groups may also apply other recent developments of testing for measurement invariance such as the alignment procedure (see, e.g., Muthén and Asparouhov, 2013) and examine the comparability of their findings to those of other more established approaches to test for invariance. Hopefully these methods and our empirical demonstration will encourage and support substantive researchers in their endeavor to conduct meaningful comparative research.

Acknowledgments

The authors would like to thank Lisa Trierweiler for the English proof of the manuscript. The work of the JC and ED was supported by the University Research Priority Program "Social Networks" of the University of Zürich.

Supplementary Material

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00733/abstract

[22]Note that when certain time points were not available we extrapolated the data.

References

Beckers, T., Siegers, P., and Kuntz, A. (2012). Congruence and performance of value concepts in social research. *Surv. Res. Method* 6, 13–24.

Beierlein, C., Davidov, E., Schmidt, P., Schwartz, S., and Rammstedt, B. (2012). Testing the discriminant validity of Schwartz' Portrait Value Questionnaire items - a replication and extension of Knoppen and Saris (2009). *Surv. Res. Method* 6, 25–36.

Billiet, J., Maddens, B., and Beerten, R. (2003). National identity and attitude toward foreigners in a multinational state: a replication. *Polit. Psychol.* 24, 241–257. doi: 10.1111/0162-895X.00327

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: Wiley. doi: 10.1002/9781118619179

Braun, M., Behr, D., and Kaczmirek, L. (2013). Assessing cross-national equivalence of measures of Xenophobia: evidence from probing in web surveys. *Int. J. Public Opin. R.* 25, 383–395. doi: 10.1093/ijpor/eds034

Brown, T. A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York, NY: Guilford Press.

Byrne, B. M., Shavelson, R. J., and Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures–the issue of partial measurement invariance. *Psychol. Bull.* 105, 456–466. doi: 10.1037/0033-2909.105.3.456

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling* 14, 464–504. doi: 10.1080/10705510701301834

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness of fit indexes to lack of measurement invariance. *Struct. Equ. Modeling* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., and Schwartz, S. H. (2014a). Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values. *Front. Psychol.* 5:982. doi: 10.3389/fpsyg.2014.00982

Cieciuch, J., Davidov, E., Vecchione, M., Beierlein, C., and Schwartz, S. H. (2014b). The cross-national invariance properties of a new scale to measure 19 basic human values a test across eight countries. *J. Cross Cult. Psychol.* 45, 764–776. doi: 10.1177/0022022114527348

Datler, G., Jagodzinski, W., and Schmidt, P. (2013). Two theories on the test bench: internal and external validity of the theories of Ronald Inglehart and Shalom Schwartz. *Soc. Sci. Res.* 42, 906–925. doi: 10.1016/j.ssresearch.2012.12.009

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

Davidov, E. (2010). Testing for comparability of human values across countries and time with the third round of the European Social Survey. *Int. J. Comp. Sociol.* 51, 171–191. doi: 10.1177/0020715210363534

Davidov, E., Cieciuch, J., Schmidt, P., Meuleman, B., and Algesheimer, R. (2015). The comparability of measurements of attitudes toward immigration in the European Social Survey: exact versus approximate measurement equivalence. *Public Opin. Qurterly* 79, 244–266. doi: 10.1093/poq/nfv008

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., and Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *J. Cross Cult. Psychol.* 43, 558–575. doi: 10.1177/0022022112438397

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., and Billiet, J. (2014a). Measurement equivalence in cross-national research. *Annu. Rev. Sociol.* 40, 55–75. doi: 10.1146/annurev-soc-071913-043137

Davidov, E., Meuleman, B., Schwartz, S. H., and Schmidt, P. (2014b). Individual values, cultural embeddedness, and anti-immigration sentiments: explaining differences in the effect of values on attitudes toward immigration across Europe. *Kölner Z. Soz. Sozpsychol.* 66, 263–285. doi: 10.1007/s11577-014-0274-5

Davidov, E., Schmidt, P., and Schwartz, S. (2008). Bringing values back in. The adequacy of the European Social Survey to measure values in 20 countries. *Public Opin. Q.* 72, 420–445. doi: 10.1093/poq/nfn035

Davidov, E., and Siegers, P. (2010). "Comparing basic human values in East and West Germany", in *Komparative Empirische Sozialforschung [Comprative Empirical Social Research]*, eds T. Beckers, K. Birkelbach, J. Hagenah, and U. Rosar (Wiesbaden: Verlag), 43–63. doi: 10.1007/978-3-531-92472-4_2

Gelman, A. (2003). A Bayesian formulation of explanatory data analysis and goodness of fit testing. *Int. Stat. Rev.* 71, 369–382. doi: 10.1111/j.1751-5823.2003.tb00203.x

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from unform. *Electron. J. Statist.* 7, 2595–2602. doi: 10.1214/13-EJS854

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. P., (eds.) et al. (2010). *Survey Methods in Multinational, Multiregional, and Multicultural Contexts.* Hoboken, NJ: John Wiley & Sons. doi: 10.1002/9780470609927

Hitlin, S., and Piliavin, J. A. (2004). Values: reviving a dormant concept. *Annu. Rev. Sociol.* 30, 359–393. doi: 10.1146/annurev.soc.30.012703.110640

Horn, J. L., and McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Exp. Aging Res.* 18, 117–144. doi: 10.1080/03610739208253916

Inglehart, R. (1977). *The Silent Revolution: Changing Values and Political Styles among Western Publics*. Princeton, NJ: Princeton University Press.

Inglehart, R., and Welzel, C. (2005). *Modernization, Cultural Change, and Democracy. The Human Development Sequence.* Cambridge: Cambridge University Press.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika* 36, 409–426. doi: 10.1007/BF02291366

Jöreskog, K. G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika* 43, 443–477. doi: 10.1007/BF02293808

Jöreskog, K. G. (1993). "Testing structural equation models," in *Testing Structural Equation Models,* eds K. A. Bollen and J. S. Long (Newbury Park, CA: Sage), 294–316.

Jouha, J., and Moustaki, I. (2013). *Non-Equivalence of Measurement in Latent Variable Modelling of Multigroup Data: a Sensitivity Analysis.* Available online at: http://ssrn.com/abstract=2332071

Jowell, R., Roberts, C., Fitzgerald, R., and Eva, G. (2007). *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey.* London: Sage.

Kaplan, D. (2014). *Bayesian Statistics for the Social Sciences*. New York, NY: Guilford Press.

Kass, R. E., and Raftery, A. E. (1995). Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795. doi: 10.1080/01621459.1995.10476572

Kruschke, J. K., Aguinis, H., and Joo, H. (2012). The time has come! Bayesian methods for data analysis in the organizational sciences. *Organ. Res. Methods* 15, 722–752. doi: 10.1177/1094428112457829

Kuntz, A., Davidov, E., Schwartz, S. H., and Schmidt, P. (2015). Human values, legal regulation and approval of homosexuality in Europe: a cross-country comparison. *Eur. J. Soc. Psychol.* 45, 120–134. doi: 10.1002/ejsp.2068

Latcheva, R. (2011). Cognitive interviewing and factor-analytic techniques: a mixed method approach to validity of survey items measuring national identity. *Qual. Quant.* 45, 1175–1199. doi: 10.1007/s11135-009-9285-0

Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Struct. Equ. Modeling* 18, 663–685. doi: 10.1080/10705511.2011.607723

Levy, R., and Choi, J. (2013). "Bayesian structural equation modeling," in *Structural Equation Modeling. A Second Course, 2nd Edn.*, eds G. R. Hancock and R. O. Mueller (Charlottesville, NC: Information Age Publishing), 563–624.

Little, T. D., Slegers, D. W., and Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Struct. Equ. Modeling* 13, 59–72. doi: 10.1207/s15328007sem1301_3

Meade, A. W., Johnson, E. C., and Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *J. Appl. Psychol.* 93, 568–592. doi: 10.1037/0021-9010.93.3.568

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58, 525–543. doi: 10.1007/BF02294825

Meuleman, B. (2012). "When are intercept differences substantively relevant in measurement invariance testing," in *Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt*, eds S. Salzborn, E. Davidov, and J. Reinecke (Wiesbaden: Springer), 97–104.

Meuleman, B., Davidov, E., Schmidt, P., and Billiet, J. (2012). "Social location and value priorities: a European-wide comparison of the relation between social-structural variables and human values," in *Society and Democracy in Europe,* eds O. Gabriel and S. I. Keil (London: Routledge), 45–67.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York, NY: Routledge.

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

Muthén, B. O., and Asparouhov, T. (2013). *BSEM Measurement Invariance Analysis. Mplus Web Notes: No. 17*. Available online at: http://www.statmodel.com

Muthén, L., and Muthén, B. O. (1998–2014). *Mplus User's Guide, 7th Edn*. Los Angeles, CA: Muthén and Muthén.

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Polit. Anal.* 22, 45–60. doi: 10.1093/pan/mpt014

Oberski, D. L., Weber, W., and Revilla, M. (2012). "The effect of individual characteristics on reports of socially desirable attitudes towards immigration," in *Methods, Theories, and Empirical Applications in the Social Sciences: Festschrift for Peter Schmidt*, eds S. Salzborn, E. Davidov, and J. Reinecke (Wiesbaden: Springer), 151–158.

Opp, K.-D. (2011). Modeling micro macro relationships: problems and solutions. *J. Math. Sociol.* 35, 209–234. doi: 10.1080/0022250X.2010.532257

Piurko, Y., Schwartz, S. H., and Davidov, E. (2011). Basic personal values and the meaning of left-right political orientations in 20 countries. *Polit. Psychol.* 32, 537–561. doi: 10.1111/j.1467-9221.2011.00828.x

Popper, K. R. (2005). *The Logic of Scientific Discovery*. London: Routledge.

Saris, W., Knoppen, D., and Schwartz, S. (2013). Operationalizing the theory of human values: balancing homogeneity of reflective items and theoretical coverage. *Surv. Res. Method* 7, 29–44.

Saris, W., Satorra, A., and van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Struct. Equ. Modeling* 16, 561–582. doi: 10.1080/10705510903203433

Sarrasin, O., Green, E. G. T., Berchtold, A., and Davidov, E. (2012). Measurement equivalence across subnational groups: an analysis of the conception of nationhood in Switzerland. *Int. J. Public Opin. R.* 25, 522–534. doi: 10.1093/ijpor/eds033

Schafer, J., and Graham, J. (2002). Missing data: our view of the state of the art. *Psychol. Methods* 7, 147–177. doi: 10.1037/1082-989X.7.2.147

Schwartz, S. H. (2003). "A proposal for measuring value orientations across nations," in *Questionnaire Development Package of the European Social Survey, Chapter 7*. Available online at: www.europeansocialsurvey.org

Schwartz, S. H., Cieciuch, J., Vecchione, M., Davidov, E., Fischer, R., Beierlein, C., et al. (2012). Refining the theory of basic individual values. *J. Pers. Soc. Psychol.* 103, 663–688. doi: 10.1037/a0029393

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* 6, 461–464. doi: 10.1214/aos/1176344136

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measure of model complexity and fit. *J. R. Statist. Soc. B.* 64, 583–639. doi: 10.1111/1467-9868.00353

Steenkamp, J.-B. E. M., and Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *J. Consum. Res.* 25, 78–90. doi: 10.1086/209528

Steinmetz, H. (2013). Analyzing observed composite differences across groups: is partial measurement invariance enough? *Methodology* 9, 1–12. doi: 10.1027/1614-2241/a000049

Thompson, M. S., and Green, S. B. (2013). "Evaluating between group differences in latent variable means," in *Structural Equation Modeling. A Second Course, 2nd Edn.*, eds G. R. Hancock and R. O. Mueller (Charlottesville, VA: Information Age Publishing), 163–218.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., and Muthén, B. O. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Front. Psychol.* 4:770. doi: 10.3389/fpsyg.2013.00770

van de Schoot, R., Lugtig, P., and Hox, J. (2012). A checklist for testing measurement invariance. *Eur. J. Dev. Psychol.* 9, 486–492. doi: 10.1080/17405629.2012.686740

Vandenberg, R. J., and Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organ. Res. Methods* 3, 4–70. doi: 10.1177/109442810031002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.