Psychological Test and Assessment Modeling, Volume 62, 2020 (2), 139-145

Editorial
Special Topic: Establishing comparability and measurement invariance in large-scale assessments, part II

Old questions, new challenges and possible solutions

Lale Khorramdel[1]*, Artur Pokropek*[2] *& Peter van Rijn*[3]

Introduction

The growing popularity of international large-scale assessments and surveys, as well as the move from paper- to digital-based administration and data collection modes, has led to new challenges and generated demand for new methodological and statistical solutions (Khorramdel, Pokropek, & van Rijn, 2020; Davidov et al., 2014; von Davier et al., 2019). When the number of groups to be analyzed is counted in the hundreds, not tens, and when new types of data like response times (Shin et al. 2020), process data, such as action sequences (von Davier et al., 2019), or text responses (Zehner et al. 2020) are collected and analyzed, new approaches are inevitable. The statistical and methodological community reacted promptly and, in the last two decades, has witnessed an unprecedented growth in models that directly address these new challenges, which have provided new insight and implications for issues of measurement invariance (MI) and the comparability of data, test scores, and measured constructs.

Multilevel analysis was employed to address the problem of comparability by considering the groups as a random mode of variation (De Jong et al., 2007; Fox, 2010). Multilevel models introduced the notion of approximate invariance by assuming that measurement parameters are approximately the same in all groups. In other words, no parame-

[1] *Correspondence concerning this article should be addressed to:* Lale Khorramdel, Ph.D., Center for Advanced Assessments (CAA), National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104, USA; email: lale.khorramdelameri@gmail.com

[2] Institute of Philosophy and Sociology of the Polish Academy of Sciences, Warsaw, Poland

[3] ETS Global, Amsterdam, The Netherlands

ters are exactly invariant across groups, but each parameter is subject to random variations causing slight differences between item parameters across groups. It was quickly noticed that the features of multilevel modeling can not only be used to establish common scales but also to test different assumptions about measurement invariance, and extensive work was conducted in this direction (Jak et al., 2013; Hartig et al., 2020).

Following the logic of multilevel modeling, multi-group Bayesian Structural Equation Modeling (MG-BSEM) approaches (Muthén & Asparouhov, 2013) were introduced. MG-BSEM models can be interpreted as restricted multilevel models where the between-group variation is forced to be small. From the multigroup modeling perspective, they could be understood as relaxed multigroup scalar models, where elastic[4] rather than fixed constraints are introduced to relax the assumption of full invariance (for an overview, see Pokropek et al., 2020).

Alignment optimization was introduced (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014), or rather reintroduced (Khorramdel, Pokropek, & van Rijn, 2020; Haberman, 2009), and gained considerable attention form researchers as an easy tool to achieve the best possible solution given the used data. This approach can be understood as attempt to reconcile the differences between approximate and partial invariance (for an overview on different types of invariance, see Davidov et al., 2014; Khorramdel, Pokropek, & van Rijn, 2020) and replaces cross-country equality constraints in multiple group models by aligning item parameters from group-specific models into a 'most optimal invariance pattern' that allows the estimation of group-specific factor means and variances.

The more traditional confirmatory factor analysis (CFA) and item response theory (IRT) models were extensively tested and adjusted to new conditions with large numbers of groups (Buchholz & Hartig, 2020), extensively used in empirical data analysis and operational analysis in international large-scale assessments, (Pokropek et. al 2017; Lee & von Davier 2020; OECD, 2013, 2017, 2020; Yamamoto, Khorramdel, & von Davier, 2013) and compared to newer approaches (Pokropek et al., 2019). These novel approaches show promising potential improvements to certain problems in MI analysis, but more research is needed.

One example of a problem related to MI analysis is the lack of established procedures for detecting non-comparable items, for example, the selection of anchor items for linking scales across groups and over time. Detecting non-comparable items in cross-country analysis is similar to detecting differential item functioning (DIF) in educational and psychological instruments (Holland & Wainer, 2012; Rutkowski et al., 2010; Swaminathan & Rogers, 1990). In classical psychological and educational settings, there is usually a limited number of groups, and sample sizes are rather small. In contrast, large scale cross-country studies deal with a large number of groups and large sample sizes (multiple countries and languages, and thousands of respondents in each country). In such settings,

[4] Elastic constraints allow for small deviations from fixed values; see Pokropek et al. (2020) for a detailed discussion.

statistical tests are oversensitive; that is, differences that show to be significant in these tests can be small and irrelevant from a practical point of view (Glas & Jehangir, 2013).

A second example of a challenge in MI analysis is the use and interpretation of model fit and item fit statistics that arise with a large number of groups that ought to be compared. Although some work has been done with regard to establishing appropriate thresholds for certain item fit statistics in the analysis of DIF (Buchholz & Hartig, 2020; Joo, et al., submitted for publication), thresholds for a broader range of fit statistics and for different data situations are not well established, forcing researchers to make more or less educated guesses (van de Vijver, 2011). Knowledge is also still limited regarding how to assess the fit of some of the new models (e.g., alignment of BSEM). Simulation studies show that most of the newly developed methods work well in terms of parameter recovery (from the true generating model) when all assumptions are met. But in real situations, we do not know which assumptions are met and which are violated.

More challenges emerge as the fields of large-scale assessments and comparative studies develop. As the number of participating countries increases and assessment technologies are being modified to accommodate the collection of more detailed and broader information, such as process and log data, new methodologies are needed to provide statistical solutions for these new data problems. As large-scale and trend assessments evolve, concerns over DIF and MI, in particular, remain at the forefront of research initiatives.

In this special issue

This special issue volume presents a collection of papers that bring new insights and propose new solutions in response to these evolving challenges and ongoing questions of large-scale assessment data. The first two papers are focused on empirical large-scale data and the establishment of comparable scales, while the other three papers present simulation studies that examine the power and properties of different methods for investigating MI. The first study deals with the impact of different administration modes and digital-based test settings on rapid guessing behavior as potential sources of MI. The second study illustrates the examination of gender DIF and potential sources of gender differences at the international level when dealing with a large number of populations and using IRT-based approaches. The third paper presents a comparison of different IRT-based linking methods (including the method used in the second paper) and considers different types of DIF situations. The fourth paper proposes a cluster method for more flexibility in selecting (groups of) anchor items for scale linking, and the fifth paper compares different loss functions for the invariance alignment method when handling the assumption of approximate MI. In the following, we will provide a more detailed summary of these papers.

Kroehne, Deribo, and Goldhammer (2020) examined and found differences of rapid guessing behavior across different administration modes and test settings in an experimental study as part of the National Educational Panel Study (NEPS). To identify rapid guessing behavior, the authors used response time measures extracted from log data and determined response time thresholds to separate solution behavior from rapid guessing

behavior. To determine adequate thresholds, different identification methods were used (normative thresholds and thresholds identified using visual inspection) and mode-specific speed differences were considered and adjusted for. The experiment revealed lower rapid guessing rates and lower rates of omitted responses in digital-based assessments with a proctored group test setting compared to paper-based assessments (with digital pens) and digital-based assessments with an un-proctored individual online test setting. Moreover, there might be some interaction effects with certain person-level variables. The study emphasizes the potential bias of differences in test-taking behavior on measurement invariance and stresses the importance of testing and accounting for such biases to maintain the comparability and interpretation of scores across administration conditions.

Khorramdel, Pokropek, Joo, Kirsch, and Halderman (2020) examined gender-specific differential item functioning (DIF) and different sources of gender differences in the PISA 2018 reading literacy scale. They illustrated the use of a multiple-group concurrent calibration based on the two-parameter logistic model (2PLM) and generalized partial credit model (GPCM) with a partial invariance assumption (used in PISA since 2015) for analyzing gender DIF and for establishing a comparable reading literacy scale across gender groups. The proposed approach allows DIF to be dealt with on the international level and with a large number of populations. Based on this new established scale, they examined the diagnostic value of the reading literacy subscales (text sources, text formats, cognitive processes) at the international level, as well as students' attitudes towards reading obtained from the student background questionnaire. Utilizing multidimensional item response theory (MIRT) models, linear regression, and other exploratory analysis, it was found that no additional value is provided by the reading literacy subscales and that gender differences might be related to reading attitudes. The presented approach also provides a guideline for examining gender DIF and gender differences in complex large-scale assessment data with various design restrictions.

Robitzsch and Lüdtke (2020) compared three different linking methods with regard to the impact of balanced and unbalanced DIF on country mean comparisons in international large-scale assessments: a concurrent calibration based on full invariance, a concurrent calibration based on partial invariance (like used in PISA since 2015), and separate calibrations with subsequent non-robust and robust linking approaches (similar to scaling procedures used in PISA until 2012). The results from a simulation study show that in the presence of biased items and balanced DIF, the full invariance and non-robust linking approaches provide (approximately) unbiased country means with similar variability. In the presence of biased items and unbalanced DIF, the partial invariance approach and the robust linking approaches outperformed the full invariance and non-robust linking approaches and strongly reduced the bias in country mean estimates. The different approaches were also illustrated using an empirical example from the PISA 2006 reading literacy data. The authors also stress that the decision of whether an item is biased or can be used as an anchor item cannot solely be based on statistical grounds but that it also needs to be considered whether the corresponding DIF effect is construct relevant or construct irrelevant.

Pohl and Schulze (2020) introduce an extension of a cluster approach for the identification of anchor items, which can be used to facilitate comparisons across groups and over time in situations where measurement invariance does not hold and partial measurement invariance should be considered. More precisely, the cluster approach (Bechger & Maris, 2015; Pohl & Schulze, 2020b), which identifies anchor items regarding intercept parameters within the one-parameter logistic test model (1PLM), was extended to intercept and slope parameters in the 2PLM. Because the proposed cluster approach yields multiple possible item clusters that may be used as anchor items, researchers are able to consider various solutions and can illustrate the uncertainty of the results due to anchor item selection. A simulation study provided strong support for the cluster algorithm's validity and ability to deal with various situations arising in DIF analysis. The approach was further illustrated using empirical data of a mathematics competence test and provides some practical guidance on choosing appropriate thresholds for determining item clusters.

Pokropek, Robitzsch, and Lüdtke (2020) examined an extension of the invariance alignment (IA) method (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2014) that can handle the assumption of approximate measurement invariance (AMI). AMI postulates the estimation of reliable and comparable parameters across multiple countries in a multiple group model without the requirement of cross-country equality constraints of parameters. More precisely, the authors discussed the generalized form of the loss function for IA proposed by Robitzsch (2019) and evaluated different forms of the loss function under different types of non-invariance situations using simulation studies and empirical data from the European Social Survey. Because the R package sirt (Robitzsch, 2019) implemented different forms of loss functions, it was used for this study and compared to the software Mplus (Muthén & Muthén, 1998-2017). Results illustrated that no significant differences exist between the two software packages in terms of parameter recovery, and that different forms of loss functions (as implemented in sirt) differ in their performance according to the recovery of group means. These findings suggest that the performance of IA heavily depends on the form of the loss function, the sample size, and the type of invariance or non-invariance, and that the loss function implemented in Mplus might not be optimal in all situations.

Acknowledgements

The work of the second author has been prepared under the project Scales Comparability in Large-Scale Cross-Country Surveys, which is funded by the Polish National Science Centre, as part of the grant competition Sonata 8 (UMO-2014/15/D/HS6/04934).

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 495−508. https://doi.org/10.1080/10705511.2014.919210

Bechger, T. M., and Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika* 80 (2), 317–40. https://doi.org/10.1007/s11336-014-9408-y

Buchholz J and J Hartig (2020). Measurement invariance testing in questionnaires: A comparison of three Multigroup-CFA and IRT-based approaches. *Psychological Test and Assessment Modeling, 62 (1)*, 29-53. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/03_Buchholz.pdf

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). *Measurement equivalence in cross-national research*. *Annual Review of Sociology, 40*, 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

De Jong, M. G., Steenkamp, J. B. E. M., & Fox, J. P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research 34 (2)*, 260-278. https://doi.org/10.1086/518532

Fox, J. P. (2010). *Bayesian Item Response Theory*. New York: Springer

Glas, C., & Jehangir, K. (2013). Modeling country-specific differential item functioning. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 97-115). Boca Raton, FL: CRC Press.

Haberman, S. J. (2009). *Linking parameter estimates derived from an item response model through separate calibrations* (Research Report RR-09-40). Princeton, NJ. Educational Testing Service. https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2009.tb02197.x

Hartig, J., Köhler, C., & Naumann, A. (2020). Using a multilevel random item Rasch model to examine item difficulty variance between random groups. *Psychological Test and Assessment Modeling*, 62 (1), 11-27. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/02_Hartig.pdf

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. New York: Routledge.

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal, 20*, 265-282. https://doi.org/10.1080/10705511.2013.769392

Joo, S.-H., Khorramdel, L., Yamamoto, K., Shin, H. J., & Robin, F. (submitted for publication, under review). Examining the impact of different item fit statistic thresholds in the IRT scaling of the PISA cognitive domains.

Khorramdel, L., Pokropek, A., & van Rijn, P. (2020). Special Topic: establishing comparability and measurement invariance in large-scale assessments, part I. *Psychological Test and Assessment Modeling, 62 (1)*, 3-10. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/01_Khorramdel.pdf

Lee, S. S., & von Davier, M. (2020). Improving measurement properties of the PISA home possessions scale through partial invariance modeling. *Psychological Test and Assessment Modeling, 62 (1)*, 55-83. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/04_Lee.pdf

OECD (2017). *PISA 2015 Technical Report,* chapter 9 (Scaling PISA Cognitive Data). OECD Publishing: Paris. http://www.oecd.org/pisa/data/2015-technical-report/09_Chapter_09_PISA2015.pdf

OECD (2020). *PISA 2018 Technical Report,* chapter 9 (Scaling PISA cognitive data). OECD Publishing: Paris. https://www.oecd.org/pisa/data/pisa2018technicalreport/Ch.09-Scaling-PISA-Data.pdf

Pohl, S., & Schulze, D. (2020b). Partial measurement invariance: Extending and evaluating the cluster approach for identifying anchor items. Manuscript submitted for publication.

Pokropek, A., Schmidt, P., & Davidov, E. (2020). Choosing priors in Bayesian measurement invariance modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2019.1703708

Pokropek, A., Borgonovi, F., & McCormick, C. (2017). On the cross-country comparability of indicators of socioeconomic resources in PISA. *Applied Measurement in Education*, *30* (4), 243-258. https://doi.org/10.1080/08957347.2017.1353985

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 26(5),* 724-744. https://doi.org/10.1080/10705511.2018.1561293

Robitzsch, A. (2019). *sirt: Supplementary item response theory models*. R package version 3.4-64. https://CRAN.R-project.org/package=sirt

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data issues in secondary analysis and reporting. *Educational Researcher, 39* (2), 142-151. https://doi.org/10.3102/0013189X10363170

Shin, H. J., Kerzabi, E., Joo, S.-H., Robin, F., & Yamamoto, K. (2020). Comparability of response time scales in PISA. *Psychological Test and Assessment Modeling, 62 (1)*, 107-135. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/06_Shin.pdf

van de Vijver, F. J. (2011). Capturing bias in structural equation modeling. In E. Davidov, P. Schmidt & J. Billiet (Eds.), *Cross-cultural analysis. Methods and applications* (pp. 3-34). New York: Routledge.

von Davier, M., Khorramdel, L., He, Q., Shin, H., & Chen, H. (2019). Developments in Psychometric Population Models for Technology-Based Large-Scale Assessments – An Overview of Challenges and Opportunities. *Journal of Educational and Behavioral Statistics, 44*, 671-705. https://doi.org/10.3102/1076998619881789

Yamamoto, K., Khorramdel, L., & von Davier, M. (2013, updated 2016). Scaling PIAAC cognitive data. In OECD (2013), Technical Report of the Survey of Adult Skills (PIAAC), chapter 17 (pp. 406-438), PIAAC, OECD Publishing. Retrieved from http://www.oecd.org/site/piaac/All%20PIACC%20Technical%20Report%20final.pdf

Zehner, F., Kroehne, U., Hahnel, C., & Goldhammer, F. (2020). PISA reading: Mode effects unveiled in short text responses. *Psychological Test and Assessment Modeling, 62 (1)*, 85-105. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/05_Zehner.pdf