Psychological Test and Assessment Modeling, Volume 62, 2020 (1), 3-10

Editorial
Special Topic: establishing comparability and measurement invariance in large-scale assessments, part I

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

Introduction

This special issue volume presents research and applications which deal with the comparability of data and test scores in the important field of large-scale assessments (LSAs). National and International LSAs like NAEP (National Assessment of Educational Progress) administered by NCES (National Center for Education Statistics), or PISA (Programme for International Student Assessment) and PIAAC (Programme for the International Assessment of Adult Competencies) administered by the OECD (Organisation for Economic Co-operation and Development) provide important data for educational and social research and are used by policymakers around the globe. While these assessments and surveys are usually low stakes for test takers, they are high stakes for countries and economies. Their results are used for numerous group-level comparisons on the national (e.g., different school types or subpopulations within a country) and international (e.g., different countries) level and for comparisons over time when measuring and reporting trends. In LSAs, group-wise comparisons are made using cognitive variables (such as math, reading or science proficiencies) as well as non-cognitive variables (such as attitudinal variables or socioeconomic variables). Moreover, relations between variables are compared across different groups to better understand the mechanism behind social and educational systems; for example, the relation between socioeconomic status and proficiency scores might be interpreted as an indicator of inequalities.

Therefore, one of the main goals in national and international large-scale surveys is to provide data and test scores that are comparable across different groups of interest and

[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: lkhorramdel@nbme.org

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

[3] ETS Global, Amsterdam, The Netherlands

over time. In the first cycle of PIAAC about 40 countries participated, and about 90 countries participated in the PISA 2018 cycle. For each of these LSAs, comparability needed to be established across countries and across languages within one assessment cycle (note that some countries test in multiple languages). Moreover, comparability needed to be established for each country over time; in case of PISA from 2000 to 2018, for example. In future assessment cycles, more countries will join these assessments, including developing countries. Without establishing comparability of the measured constructs and the underlying scales, group-level comparisons may result in seriously biased inferences and conclusions. The first step of achieving such comparability, also called measurement invariance (MI), is careful item development and test design. Comparable indicators of latent traits can only be achieved if the items or tasks used to measure the underlying construct of interest have similar characteristics across the tested populations (such as item difficulties and item discriminations) and if they measure the same latent trait.

In cognitive assessment, possible violations of MI at the item and construct level can occur in the form of differential item functioning (DIF); for example, when items show different difficulties across groups due to different cognitive demands based on language or cultural differences. Test takers from different cultures are not necessarily equally exposed to the content described in the item or item stem, or the text of an item might not be comparable across different languages with regard to structure and grammar. MI violations can also originate from technical difficulties, such as translation or scoring errors, or if there are differences in the test administration (e.g., providing different instructions to test takers) or different modes of administration (computer versus paper-based testing).

Similar issues apply to non-cognitive scales. The meaning of different items may differ across countries and across time, resulting in DIF. A good example is the PISA home possessions scale which is measured with three types of items: items referring to the general wealth of the students' family (possession of own room, internet, dishwasher, cellular phones, televisions, computers, cars), items addressing cultural possessions (literature, poetry, art, books at home) and items indicating educational resources (desk, study place, software, textbooks, dictionary). Potential problems with different meanings of items across groups in relation to the home-possession scale seem obvious. For example, having a car may not indicate socio-economic status the same way in the United States, where almost everyone has a car independent of the level of income (due to larger distances between locations and limited public transportation), as in Japan, where car ownership is less common, even in relatively wealthy families (distances are shorter between locations and public transportation is widespread and efficient). The same logic applies to changes within country and over time in the prevalence of certain possessions. For instance, between 2000 and 2018 there was a significant increase in the availability of mobile phones in OECD countries (Pokropek, Borgonovi, & McCormick, 2017; Lee & von Davier, 2020, this issue).

Violations of MI might still occur to some extent after careful item and instrument development, well-planned test designs, and careful test administration, and, therefore, have to be accounted for when analyzing the data in the second step. Statistical methods

such as linking and psychometric modeling are used to examine measurement invariance (MI) and to account for its violations to ensure valid results can be provided.

Differential item functioning and measurement invariance

Modern empirical work on measurement invariance and comparability has emerged from two traditions: the first tradition comes from educational and psychological assessments, and the second tradition comes from cross-cultural studies and large-scale cross-country surveys.

Researchers coming from the field of educational and psychological measurement were interested in detecting "problematic items" that show DIF and, hence, could be biased against certain subgroups. A series of methods based on observed scores was developed to fulfill this goal. Among the most prominent is the *delta-plot* method from Angoff (1972), adopted from the Mantel–Haenszel procedure, which originates from epidemiology (Mantel & Haenszel, 1959; for an overview see Holland & Wainer, 2012; Swaminathan & Rogers, 1990). Much later, latent variable models in the form of item response theory (IRT) models were used for DIF detection (Thissen, Steinberg, & Wainer, 1993). DIF detection methods were used for purification purposes with items showing the greatest DIF effects being eliminated from the item pool in an iterative process (Lord, 1980). Although it is impossible to achieve full item purification (Thissen et al., 1993) such procedures were designed to obtain a scale which was free of large biases and could be prepared for other statistical treatments like test equating.

Researchers coming from the cross-cultural tradition focused on measurement invariance in non-cognitive measures using the latent variable framework. The multiple-group confirmatory factor analysis (CFA) modeling approach, which is used for this purpose, was introduced by Jöreskog (1971), who, interestingly, illustrated this idea using data from cognitive tests directly referring to the theory of mental test scores from Lord and Novick (1968). Cross-cultural researchers were more interested in describing the global characteristics of the scales. Their framework describes different types of measurement invariance. If the measurement model used for different countries has the same structure, the measured scale possesses *configurational invariance*; an assumption which was being tested using exploratory factor analysis (EFA). Configurational invariance is the weakest form of invariance and determines whether respondents in different countries use the same conceptual framework when answering particular survey questions while item parameters can differ (Cheung & Rensvold, 2002; Horn & McArdle, 1992; Vandenberg & Lance, 2000). This is a necessary, but not a sufficient condition for a scale to be comparable across different groups. It does not allow us to draw valid comparisons of relations between the analyzed scale and other variables. Such comparisons can be conducted only in the presence of *metric invariance*, also called weak factorial invariance. This type of invariance applies when the loadings of indicators on the factors are equal across respondents in different groups, and has usually been tested by multiple group CFA models with different sets of constraints. However, even in the presence of configural and metric invariance, full comparability of scales across groups cannot be achieved.

According to the CFA framework, this can only be achieved in the presence of *scalar invariance*, also called strong factorial invariance. Scalar invariance requires that the factor loadings and the intercepts for each item are the same across groups (for an overview, see Davidov et al., 2014). Another level of invariance *is partial invariance*, also called local misspecification, where some of the items are fully invariant while others are non-invariant. While this type of invariance was discussed in the CFA framework (Byrne et al., 1989) much less attention was given to it (for an overview see Pokropek et al., 2019)..

Context of large-scale assessments

In the field of large-scale assessments, both MI traditions and the respective statistical approaches for investigating MI are converging. Large-scale assessments can contain both cognitive tests and questionnaires or surveys measuring noncognitive constructs, and experts from both fields discuss their methods and work on joint approaches. It often eventuates that both sides have similar ideas but utilize different terminologies. One example is the concept of measurement alignment, introduced for CFA models by Asparouhov and Muthén (2014) and subsequently for IRT models (Muthén & Asparouhov, 2014). In the context of linking educational tests using IRT, a similar approach was proposed by Haberman (2009) and referred to as a simultaneous linking approach (see the comment by von Davier in Avvisati, Donné, Paccagnella, 2019). Large-scale assessment methodologies heavily rely on IRT-based approaches and IRT modeling is not only used for scaling the items of cognitive domains, but for noncognitive constructs, too.

However, learning from both traditions and utilizing their statistical modeling approaches can lead to a deeper understanding of different MI issues and lead to better and more efficient methodologies for investigating and treating these issues. This is especially important as new challenges emerge for data analysis and MI with the introduction of computer-based administration modes and more countries and economies joining international large-scale surveys. While the inclusion of developing countries, in addition to OECD and developed countries (like it will be the case for PISA 2021), will lead to a higher diversity in DIF and the range of test scores, the move from paper-based to computer-based assessments (like in PISA 2015) is leading to more diversity in the data as new item types are introduced (e.g., simulation-based items) and log-file or process data (e.g, action and timing data) are available. Hence, problems of comparability and MI are no longer restricted to cognitive and noncognitive items but may need to be accounted for in process data as well.

In this special issue

This special issue volume will provide an overview of some of the most recent statistic and psychometric approaches for examining DIF and MI based on factor analysis, generalized linear models, and IRT. The presented studies are based on simulation studies and empirical applications, utilize methods from natural language processing (NLP), and deal

with the comparability of cognitive and noncognitive constructs across different groups within one assessment cycle and across different cycles over time, the impact of different item types on MI, the comparability of data across different administration modes when moving from a paper-based to a computer-based assessment, and the comparability of response time data across multiple populations.

Hartig, Köhler, and Naumann (2020) propose three modeling approaches to test for random group differential item functioning (RG-DIF): (1) three-level Generalized Linear Mixed Models (GLMMs), (2) three-level GLMMs with anchor items, and (3) item-wise multilevel logistic regression (ML-LR). More precisely, these models use variances of item difficulties at the group level to test for violations of MI of item parameters across these groups. The performance of the models was investigated through a simulation study and results show that all three methods performed well in the case of uncorrelated RG-DIF. However, in the case of correlated RG-DIF, estimated variances were biased for the full three-level GLMM and ML-LR while the three-level GLMM with anchor items allowed unbiased estimation of RG-DIF. Anchor items seem to offer a solution but need to be unbiased by RG-DIF which is not necessarily the case in empirical data.

Buchholz and Hartig (2020) compare existing approaches for testing MI in questionnaire data including multigroup confirmatory factor analysis (MGCFA) for ordinal and continuous data and the multigroup IRT-based approach introduced in PISA 2015. All three approaches were applied to simulated data containing different types and extents of MI violations, and to empirical data from the PISA 2015 student questionnaire. Violations to MI were identified using indices of the *magnitude* and *direction* of local parameter misfit (modification indices and SEPC in the MGCFA, RMSD and MD in the IRT approach). Results indicate that measures from all three approaches were consistent in identifying group differences in item difficulty parameters. With regard to group differences in the item slope parameter, the IRT approach was able to identify both negative and positive deviations from the true parameter, while the MGCFA identified mostly negative deviations. The paper discusses the differences of the used fit statistics for identifying MI violations and their different advantages.

Lee and von Davier (2020) examine the longitudinal and cross-country MI of the PISA home-possessions scale, one of the three components used to measure socioeconomic status (SES). Similar to Buchholz and Hartig (2020), they used the IRT scaling approach which was introduced in PISA 2015, applied it to a large data set combining PISA data from different assessment cycles, and (like in PISA 2015) allow for partial invariance. More precisely, common item parameters were estimated across different groups for most items while group-specific item parameters were estimated in case of item misfit (i.e. group-specific misfit to the common item parameter estimated across groups). Results show that most of the items in the scale are invariant over time but not invariant across countries. Moreover, it was found that the PISA 2015 scaling approach outperformed past PISA scaling approaches by achieving a higher comparability of the home-possessions scale over time and across countries, and in some cases even improved the within-country accuracy of the home-possessions scores.

Zehner, Kroehne, Hahnel, and Goldhammer (2020) examine the comparability of short text responses in the PISA reading scale across a paper-based and computer-based presentation mode. Data were gathered in an experimental setting, in which a student cohort from Germany answered PISA 2012 reading items on either computer, paper, or both. Text response features were extracted using natural language processing techniques with regard to the quantity (proposition entity count) and quality (relevance proportion) of information. The study found that students incorporated more information into their text responses on computer than on paper, with some items being more affected than others. More precisely, correct responses in the computer-based assessment tended to contain more pieces of information and larger proportions of relevant information. Moreover, item seemed to be slightly harder on the computer even though the number of omitted responses on the computer was lower. Results also hint towards a possible relationship between these mode effects and gender. The paper discusses possible sources of mode effects and stresses to account for them when interpreting trend measures across different administration modes.

Shin, Kerzabi, Joo, Robin, and Yamamoto (2020) investigate the possibility of generating response time (RT) scales in PISA that are comparable across countries for their utilization in data quality assurance and data analysis. This feasibility study uses RT data from PISA 2015 where a computer-based assessment was implemented as the main administration mode for the majority of participating countries. To properly deal with RT outliers and large sets of missing data, like introduced by the PISA test design, RTs were categorized at the item-level in a specific way. The resulting RT data were then analyzed using unidimensional and multidimensional multiple-group item response theory (IRT) models. Results show a better fit of the multidimensional model indicating that RTs vary by item type (multiple-choice versus constructed-response items) which has to be accounted for when establishing comparable RT scales. The study discusses different options for categorizing RT data and the problem of interpreting the IRT-based intercept parameter for RTs in the presented approach. Moreover, implications for the analytical procedures involving RT in international large-scale assessments are provided.

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

Angoff, W. H. (1972) *A technique for the investigation of cultural differences*. Paper presented at the annual meeting of the American Psychological Association, Honolulu, May 1972.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508. DOI: 10.1080/10705511.2014.919210

Avvisati, F., Le Donné, N., & Paccagnella, M. (2019). A meeting report: Cross-cultural comparability of questionnaire measures in large-scale international surveys. *Measurement Instruments for the Social Sciences* 1, 8 (2019). DOI:10.1186/s42409-019-0010-z

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456. DOI: https://doi.org/10.1037/0033-2909.105.3.456

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255. DOI: https://doi.org/10.1207/S15328007SEM0902_5

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

Haberman, S. J. (2009). Linking parameter estimates derived from an item response model through separate calibrations. *ETS Research Report ETS RR-09-40*. Princeton: ETS. https://files.eric.ed.gov/fulltext/EJ1110994.pdf

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

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117-144. DOI: https://doi.org/10.1080/03610739208253916

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409-426. DOI: ttps://doi.org/10.1007/BF02291366

Lance, C. E., Vandenberg, R. J., & Self, R. M. (2000). Latent growth models of individual change: The case of newcomer adjustment. *Organizational Behavior and Human Decision Processes*, *83*(1), 107-140. DOI: https://doi.org/10.1006/obhd.2000.2904

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Lawrence Erlbaum.

Lord and Novick (1968) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748. DOI: https://doi.org/10.1093/jnci/22.4.719

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology | Quantitative Psychology and Measurement*, *5*:978. DOI: 10.3389/fpsyg.2014.00978

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. DOI: 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. DOI: 10.1080/10705511.2018.1561293

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370. DOI: https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L. & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland and H. Wainer (ed.), *Differential Item Functioning* (pp. 67–115). Hillsdale: Lawrence Earlbaum.