OECD Education Working Papers No. 201

Invariance analyses in large-scale studies

Fons J. R. Van de Vijver,
Francesco Avvisati,
Eldad Davidov,
Michael Eid,
Jean-Paul Fox,
Noémie Le Donné,
Kimberley Lek,
Bart Meuleman,
Marco Paccagnella,
Rens van de Schoot

https://dx.doi.org/10.1787/254738dd-en

OECD

Organisation for Economic Co-operation and Development

EDU/WKP(2019)9

Unclassified **English text only**

30 April 2019

DIRECTORATE FOR EDUCATION AND SKILLS

INVARIANCE ANALYSES IN LARGE-SCALE STUDIES

OECD Education Working Paper No. 201

Fons J. R. van de Vijver (Tilburg University, Editor); Francesco Avvisati (OECD); Eldad Davidov (University of Cologne and University of Zurich); Michael Eid (Free University of Berlin); Jean-Paul Fox (University of Twente); Noémie Le Donné (OECD); Kimberley Lek (Utrecth University); Bart Meuleman (KU Leuven); Marco Paccagnella (OECD); and Rens van de Schoot (Utrecht University)

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Francesco Avvisati (francesco.avvisati@oecd.org); Noémie Le Donné (noemie.ledonne@oecd.org); and Marco Paccagnella (marco.paccagnella@oecd.org)

JT03446903

OECD Education Working Papers Series

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

Acknowledgements

This report is the result of the work conducted by a group of experts, contracted by the OECD to provide advice and guidance on how to address the issue of measurement invariance in international surveys. Francesco Avvisati, Miloš Kankaraš, Noémie Le Donné and Marco Paccagnella followed and co-ordinated the work of the group. Chapter 1 has been drafted by Fons van de Vijver, who chaired and supervised the work of the group. Chapter 2 has been drafted by Eldad Davidov and Bart Meuleman. The authors of Chapter 3 are Kimberley Lek and Rens van de Schoot. Jean-Paul Fox authored Chapter 4, and Michael Eid wrote Chapter 5. Chapter 6, drafted by Francesco Avvisati, Noémie Le Donné and Marco Paccagnella, reports on the presentations and discussions held in the course of a conference organised at the OECD in November 2018 to discuss issues surrounding measurement invariance in international surveys. Florence Bernard and Hanna Varkki provided administrative and logistical support in the organisation of the conference. Hanna Varkki provided editorial assistance in the publication of this working paper.

Abstract

Large-scale surveys such as the Programme for International Student Assessment (PISA), the Teaching and Learning International Survey (TALIS), and the Programme for the International Assessment of Adult Competences (PIAAC) use advanced statistical models to estimate scores of latent traits from multiple observed responses. The comparison of such estimated scores across different groups of respondents is valid to the extent that the same set of estimated parameters holds in each group surveyed. This issue of invariance of parameter estimates is addressed in model fit indices which gauge the likelihood that one set of parameters can be used across all groups. Therefore, the problem of scale invariance across groups of respondents can typically be framed as the question of how well a single model fits the responses of all groups. However, the procedures used to evaluate the fit of these models pose a series of theoretical and practical problems. The most commonly applied procedures to establish invariance of cognitive and non-cognitive scales across countries in large-scale surveys are developed within the framework of confirmatory factor analysis and item response theory. The criteria that are commonly applied to evaluate the fit of such models, such as the decrement of the Comparative Fit Index in confirmatory factor analysis, work normally well in the comparison of a small number of countries or groups, but can perform poorly in large-scale surveys featuring a large number of countries. More specifically, the common criteria often result in the non-rejection of metric invariance; however, the step from metric invariance (i.e. identical factor loadings across countries) to scalar invariance (i.e. identical intercepts, in addition to identical factor loadings) appears to set overly restrictive standards for scalar invariance (i.e. identical intercepts). This report sets out to identify and apply novel procedures to evaluate model fit across a large number of groups, or novel scaling models that are more likely to pass common model fit criteria.

Using both real and simulated data, the following procedures are described and applied: multigroup confirmatory factor analysis, followed by alignment analysis of the same data set; Bayesian approximate measurement invariance; Bayesian measurement invariance testing in Item-Response Theory (IRT) models; and multigroup and multilevel latent class analysis. These approaches have the potential to resolve recurrent fit problems in invariance testing. Though promising, more work with these new approaches is needed to establish their suitability in large-scale surveys. The last chapter reports the conclusions from a conference in which these approaches were discussed, along with traditional approaches, in order to provide recommendations for how to address invariance issues in OECD education surveys.

Résumé

Les enquêtes à grande échelle comme le Programme international pour le suivi des acquis des élèves (PISA), l'Enquête internationale sur les enseignants et l'apprentissage (TALIS) et le Programme d'évaluation internationale des compétences des adultes (PIAAC) utilisent des modèles statistiques avancés pour produire des estimations des scores des traits latents à partir de multiples réponses observées. Une partie importante de l'analyse consiste à examiner si le même ensemble de paramètres estimés s'applique à chaque groupe étudié. Cette question de l'invariance des estimations des paramètres est abordée par les indices d'ajustement du modèle qui évaluent la probabilité qu'un ensemble de paramètres puisse être utilisé dans tous les groupes. Par conséquent, le problème de l'invariance d'échelle entre les groupes de répondants peut généralement être formulé comme la question de savoir dans quelle mesure un modèle unique correspond aux réponses de tous les groupes. Toutefois, les procédures utilisées pour évaluer l'adéquation de ces modèles posent une série de problèmes théoriques et pratiques. Les procédures les plus couramment appliquées pour établir l'invariance des échelles cognitives et non cognitives entre les pays dans les enquêtes à grande échelle sont élaborées dans le cadre de l'analyse factorielle confirmatoire et de la théorie des réponses aux items. Les critères couramment appliqués pour évaluer l'adéquation de tels modèles, tels que la diminution de l'indice d'adéquation comparative dans l'analyse factorielle confirmatoire, qui fonctionnent bien dans la comparaison d'un petit nombre de pays ne fonctionnent pas bien en pratique dans les applications à grande échelle. Plus précisément, les critères communs aboutissent souvent au non-rejet de l'invariance métrique; cependant, le passage de l'invariance métrique (c.-à-d. des coefficients de saturation identiques d'un pays à l'autre) à l'invariance scalaire (c.-à-d. des constantes identiques, en plus des coefficients de saturation identiques) semble établir des normes trop restrictives pour l'invariance scalaire (c.-à-d. des constantes identiques). La présente étude a pour but d'identifier et d'appliquer de nouvelles procédures pour évaluer l'ajustement du modèle pour un grand nombre de groupes, ou de nouveaux modèles de mise à l'échelle qui sont plus susceptibles de satisfaire aux critères communs d'ajustement du modèle.

En utilisant des données simulées et des données réelles, les procédures suivantes sont décrites et appliquées: analyse factorielle confirmatoire multigroupe, suivie d'une analyse d'alignement de la même base de données; analyse bayésienne d'invariance approximative des mesures; test bayésien d'invariance des mesures dans les modèles de réponse à l'item; et analyse de classe latente multigroupe et multiniveau. Ces approches ont le potentiel de résoudre les problèmes d'ajustement récurrents dans les tests d'invariance. Bien que prometteuses, ces nouvelles approches doivent faire l'objet d'un travail plus poussé pour établir leur pertinence dans le cadre d'enquêtes à grande échelle. Le dernier chapitre présente les conclusions d'une conférence de l'OCDE où ces approches ont été discutées en même temps que les approches traditionnelles, afin de fournir des recommandations sur la manière de traiter les questions d'invariance dans les enquêtes de l'OCDE sur l'éducation.

Table of contents

Tables

Figures

Chapter 1. Introduction

Fons J.R. van de Vijver

Why This Report?

Large-scale surveys are coming of age. In an era of globalisation, surveys that involve multiple countries have become available. A good example is the Programme for International Student Assessment (PISA). Since its first wave in 2000, PISA has grown in size from 28 countries to well over 70 countries. Information about educational systems in other countries and the comparisons of scores in "league tables" have become important benchmark information for policy makers in participating countries. However, such comparisons of scales across countries are beset with important methodological challenges. This report addresses what is often viewed as the major methodological challenge of large-scale surveys: the assessment of comparability of constructs and data.

If a mathematics test is administered in multiple countries and the aim is to compare performance across countries, it is incumbent on the team conducting the study to demonstrate that the instrument is adequate in all countries and that scores can be compared across countries. Similarly, in studies comparing the well-being or attitudes towards immigrants of respondents across multiple countries, some proof must be provided that the responses are comparable and that the models from which these scales are built apply in all countries. In the past various procedures have been proposed to test for the equivalence of instruments across countries. In particular, two types of procedures have become very common to assess equivalence (invariance) in large-scale assessment: procedures based on confirmatory factor analysis (often used for the background questionnaires assessing non-cognitive variables such as attitudes, motivation and interests) and procedures based on item response theory (often used for educational achievement data). This report mainly addresses these two types of approaches.

Both types of approaches share a common problem: There is no single, widely accepted procedure that can adequately analyse whether scores are comparable across all participating countries. Existing procedures often work well in comparisons of a few countries (in the sense that they provide estimates for all relevant parameters, such as factor loadings and item difficult estimates, combined with fit indices that provide useful information about the suitability of the model with invariant parameters), but fall short in large-scale applications.

The statistical problem of testing a measurement invariance assumption for two groups (or a small number of groups) is different from testing this assumption for many groups. The common statistical tests for measurement invariance are meant to compare two groups, which leads to evaluating two competing models (i.e. two hypotheses). The tests are also used to compare multiple groups, but this leads to simultaneously testing multiple hypotheses, which is much more complicated than evaluating a single hypothesis. For multiple hypotheses, the Type-1 errors will be substantially higher, when following the same rejection rules as for a single hypothesis. Although a Bonferroni correction can be applied to control for the inflated Type-1 error rate, the stepwise procedure of testing multiple hypotheses will influence the results, and there is a high risk of errors due to chance capitalisation. Each measurement invariance hypothesis is composite, which leads to more complex dependencies between the multiple hypotheses. Furthermore, the test statistics of the multiple hypotheses are correlated, but are treated as independent, which

can lead to biased test results. Finally, while controlling the Type-1 error, the false negative rate (i.e. the proportion of false negatives) can still be unacceptable. In practice, to avoid evaluating the entire set of hypotheses, a null model is compared to a set of restricted alternative models (already excluding multiple hypotheses). However, this restricted set might not include the optimal model, and this can lead to inferior test results.

Invariance analyses are based on assumptions about the design and data analysis that may not apply. Examples dealing with design features involve features of the instrument, such as the complete translatability and full linguistic comparability of all stimulus materials. Assumptions could also refer to the data, such as assumptions about data distributions. Given these restrictive assumptions, it should come as no surprise that fit indices often indicate a poor fit. A recurrent problem is that fit indices suggest that data cannot be compared, despite the tremendous effort that typically went in their development. Furthermore, the reasons for the poor fit are usually hard to understand: Is the poor fit a consequence of a model misspecification (and should the model that parameters are invariant across groups be rejected) or highly sensitive fit indices (that flag non-invariance while the differences in parameters across groups are very small)? This report explores novel approaches to invariance that have the potential to overcome at least some of the limitations of extant approaches.

The report is meant for researchers and students working with the international data sets. The report describes issues of current approaches and highlights promising areas to advance the field of invariance testing.

Terminology and Outline

The seminal work by Jöreskog (1969[1]; 1971[2]) on structural equation models and by Rasch (1960[3]) on item response theory have provided a major impetus to the examination of identity of model parameters across populations. Statistically rigorous tests for whether item characteristics, such as their factor loadings or difficulties, were identical across populations, became available. Rather than describing the history since the original publications, the emphasis here is on the current state of affairs in statistical models used in large-scale surveys.

In what could be called the first wave of invariance testing, the emphasis was on *exact* approaches. The statistical procedures tested the null hypothesis that some set of model parameters (factor loadings and intercepts as most important examples in structural equation models and item discrimination and item difficulties as their counterparts in item response theory) is identical across groups. These approaches are called exact because the hypothesis of interest is identity of parameters across groups (this characteristic of exact identity of parameters is released in approximate Bayesian approaches described below). Three (increasingly restrictive) types of identity are commonly assessed: *configural, metric* and *scalar invariance*.

Using the terminology of structural equation modelling, configural invariance means that a set of items shows the same pattern of salient loadings on a construct (for the simplicity of presentation, we assume here that the underlying construct is unidimensional, as many large-scale surveys assess constructs using unidimensional scales). Metric invariance means that the factor loadings are identical across groups. We then do not only know that the latent construct is comparable across groups (implied by configural invariance), but we also know that the association between items and the underlying construct is identical across groups. Any given item is an equally adequate indicator of the construct in each country. The highest level of invariance, required to compare scores across groups, is

achieved when the regression line that links the latent construct to the item scores has both the same slope (i.e. the same factor loading as required for metric invariance) and the same intercept. If the latter is not the case, an item is said to be biased or showing differential item functioning (DIF) (Holland and Wainer, 1993[4]; Van de Vijver and Leung, 1997[5]).

Since higher levels of invariance are more restrictive, these are more difficult to obtain. Ample experience with conducting tests of the three types of invariance in large-scale surveys has shown that scalar invariance is often rejected for scales in multigroup confirmatory factor analysis (described in more detail in Chapter 2).

In the first part of Chapter 2, the approach is illustrated in a new, yet increasingly important context: to establish cross-wave stability of item parameters. There is an increasing number of large-scale surveys that have multiple waves (such as PISA and the European Social Survey). Many surveys have a core of instruments that is administered in each wave. The question then arises to what extent item parameters remain identical across waves: Do constructs start to change in meaning across time or do groups change their endorsement of the construct across time? As illustrated in Chapter 2, multigroup confirmatory factor analysis is suitable to address these questions.

Historically, the first attempt to deal with the problem of poor fit in multigroup confirmatory factor analysis is due to Byrne, Shavelson and Muthén (1989[6]). Their partial measurement invariance approach releases the factor loadings and/or intercepts of designated items (based on conceptual grounds or fit statistics, either all at once or one by one) while the other parameters of the other items of the scale are kept invariant across groups. The approach may work well in small-scale applications but does not provide a viable approach in large-scale surveys where often most, if not all items have to be released.

The other approaches to deal with measurement invariance in the case of unidimensional, continuous traits that are described in the present report abandon the idea of exact invariance and start from models that allow some wiggle room in parameters, which may make these more realistic than what is done in the exact invariance approach; exact invariance is replaced here by *approximate invariance*. The first, called alignment (described in the second part of Chapter 2), is a two-step procedure in which in the first step a configural model is identified that represents the best-fitting model among all multigroup factor analytic models. In the second step this configural model undergoes an optimisation process such that for every group factor mean and variance parameter, factor loadings and intercept parameters are estimated with the same likelihood as the configural model. The factor mean and factor variance are chosen in such a manner that the total amount of measurement invariance is maximised. This approach can be evaluated using both a frequentist (i.e. maximum likelihood: ML) and a Bayesian approach. Another closely related approach, Bayesian structural equation modelling (BSEM), allows parameters to vary somewhat across groups (a chosen prior distribution defines the extent to which variation is allowed). So, loadings and intercepts are approximately identical across countries. The procedure can be combined with alignment, as illustrated in Chapter 3.

A novel Bayesian approach is described in Chapter 4. In contrast to the Bayesian random parameter approaches, which allow variation in parameters so that persons can be measured on a common latent scale in the presence of items that are not measurement invariant, this approach focuses on assessing measurement invariance, by quantifying the evidence in favour of non-invariance against the evidence in favour hypothesis of full measurement invariance. The procedure is presented in the context of the one-parameter Item-Response Theory (IRT) model and requires only estimating a marginal model (where all types of

invariance violations contribute to the covariance matrix of error terms). The Bayesian hypothesis testing approach does not rely on asymptotic results (i.e. asymptotic sampling distribution of the test statistic), and can take all sources of uncertainty into account (i.e. does not rely on parameter estimates). The Bayesian testing/marginal modelling approach is designed to identify which of the items (or sets of items) are measurement invariant and which are not.

An approach to deal with the measurement of non-ordered, categorical traits is given in Chapter 5. This chapter shows, within the framework of exact invariance testing, how multilevel and multigroup *latent class analysis* can be used to establish the existence of common and unique classes of individuals across all groups (countries) that participate in a survey. Chapter 5 describes the procedures and illustrates them on a set of TALIS measures dealing with distributed leadership in schools.

This report gives an overview of novel invariance approaches; yet, there is no attempt to provide a comprehensive overview. An approach that is not discussed here is exploratory structural equation modelling (Asparouhov and Muthén, 2009[7]). This procedure is suitable for multifactorial models by allowing some non-zero loadings of items on non-target factors. BSEM is an alternative to this procedure. Exploratory structural equation modelling is not discussed here as relatively few scales in large-scale surveys are multifactorial.

Conclusion

Examining invariance in large-scale studies continues to be problematic. Various procedures have been proposed and have shown problems.

In the present report we have gone beyond the conventional multigroup confirmatory factor analysis (MGCFA) and IRT methods by describing and applying novel approaches to scaling response data and testing model invariance, notably alignment (used with maximum likelihood or Bayesian estimation), Bayesian approximate invariance testing, Bayesian marginal invariance testing, and latent class modelling. The following four chapters demonstrate the potential of each of the procedures. However, it should be emphasised that these demonstrations are mainly a "proof of concept" and do not yet provide a decisive answer as to whether their application would mitigate or eliminate extant problems with the conventional MGCFA and IRT approaches. More experience is needed before we can decide that these approaches can live up to the expectations.

The overview of procedures in this report is not exhaustive. Thus, the report does not discuss the "old" approach of using exploratory factor analysis, followed by target rotations (Van de Vijver and Leung, 1997[5]), nor exploratory structural equation modelling (ESEM) (Asparouhov and Muthén, 2009[7]).

The field of invariance testing has undergone a major metamorphosis in the last decades. It can be expected that the field will continue to develop. Important developments could be further technical refinements in new procedures that can increasingly deal with the intricacies of large-scale cross-cultural comparisons as well as more empirical demonstrations of procedures that (do not) work well. The riddle of how to compare data across cultures in huge studies is not yet resolved.

Chapter 2. Measurement Invariance Analysis using Multiple Group Confirmatory Factor Analysis and Alignment Optimisation

Eldad Davidov and Bart Meuleman

Multiple Group Confirmatory Factor Analysis (MGCFA)

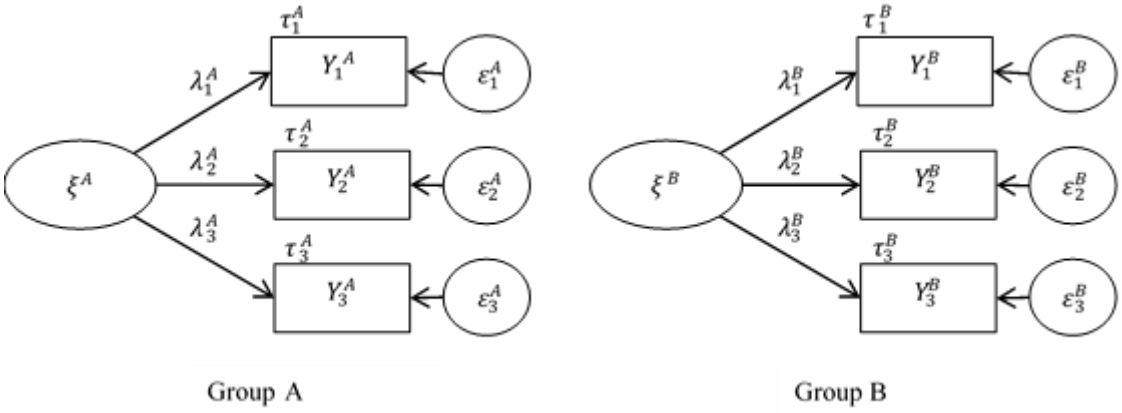
In the relevant literature, various ways to test for measurement invariance have been proposed, differing in the assumed measurement level of indicators, conceptualisation of latent variables (continuous vs. categorical) and the link function between indicators and latent variables (Meredith, 1993[8]; Davidov et al., 2014[9]). One of the most often used techniques is multigroup confirmatory factor analysis (MGCFA) – recently also called the exact measurement invariance approach as opposed to the approximate (Bayesian) approach (see the Chapter 4 in this report). MGCFA assumes a linear function between the metric indicator variables and continuous latent variables. However, this approach has been used commonly with Likert scales (which are, strictly speaking, ordinal rather than metric) when sample sizes are rather large like in the Programme for International Student Assessment (PISA) or the European Social Survey (ESS). MGCFA assumes a population divided in subgroups g, and estimates a measurement model per group. Concretely, the response on indicator variable y_i is modelled as a function of one or more latent variables ξ_j. τ_i is the intercept of this function, and factor loading λ_{ij} expresses the strength of the relationship between latent variable ξ_j and indicator y_i. Note that the measurement parameters (intercepts and factor loadings) in this model can be different across groups:

$$y_i^g = \tau_i^g + \lambda_{ij}^g \xi_j^g + \varepsilon_i^g \qquad \textbf{Equation 2.1}$$

Figure 2.1 provides a graphical illustration of the MGCFA model with two groups and a single latent variable.

Figure 2.1. Graphical representation of a MGCFA model

A latent concept (ξ) is measured by three indicators (Y1 – Y3) across two groups (A and B)



Legend: ξ=latent variable; λ=factor loading; τ=intercept; X=indicator; ε=measurement error

In the MGCFA approach, measurement invariance is tested by assessing to what extent the measurement models are similar across groups. MGCFA differentiates between three levels of invariance: configural, metric and scalar. These levels are hierarchical: Higher levels impose more restrictions on the measurement parameters, but at the same time allow a higher degree of comparability. Configural invariance requires that factor structures are equal across groups, i.e. that the same items are used to measure the same latent variables. In other words, the different groups are expected to exhibit identical patterns of salient and non-salient factor loadings. Formally, this can be written as follows:

- if λ_{ij}^{g} is close to 0, then λ_{ij}^{h} is close to 0 for *g,h = 1...G* (where superscripts *g* and *h* refer to two different groups)
- if λ_{ij}^{g} is not close to 0, then λ_{ij}^{h} is not close to 0 for *g,h = 1...G; g≠h*

Metric invariance requires in addition that the factor loadings are equal across groups:

$$\lambda_{ij}^{g} = \lambda_{ij}^{h} \text{ for } g,h = 1 \ldots G \qquad \textbf{Equation 2.2}$$

Scalar invariance furthermore requires that the items' intercepts are equal:

$$\tau_{j}^{g} = \tau_{j}^{h} \text{ for } g,h = 1 \ldots G \qquad \textbf{Equation 2.3}$$

Whereas configural invariance does not allow any comparisons of scores across groups, metric invariance guarantees the comparability of parameters expressing the relationships between concepts (such as covariances or unstandardised regression effects). Scalar invariance is a necessary condition to make valid comparisons of latent means.

MGCFA invariance testing typically begins with single group confirmatory factor analyses (CFAs) to examine whether the model fits well the data in each of the groups. If that is the case, one continues with a multigroup test of the configural invariance (i.e. equal factor structures but no equality constraints on the parameters). If the condition of configural invariance is fulfilled, one adds cross-group equality constraints on the factor loadings

(metric invariance) and subsequently on the intercepts (scalar invariance) (Steenkamp and Baumgartner, 1998[10]; Vandenberg and Lance, 2000[11]).

In order to evaluate whether measurement invariance is given, one can rely on several global fit measures produced by the statistical software. One approach consists of performing chi-square difference tests to evaluate which level of equivalence fits the data best. A major limitation of this approach, however, is that chi-square tests are known to be overly sensitive: Even substantially irrelevant differences between groups can turn up as statistically significant, especially when sample sizes are large and when the data is not normally distributed (Saris, Satorra and van der Veld, 2009[12]). A related approach consists of inspecting modification indices. For each parameter constraint a modification index is estimated, indicating by how much the chi-square value of the model would improve if that particular constraint were removed. As such, modification indices are chi-square-test statistics (with one degree of freedom) for the constrained parameter. Significant modification indices are indicative of model misfit, and the parameter constraints they refer to represent misspecifications. However, also here minor misspecification can lead to highly significant modification indices, particularly when the sample size is large, thereby limiting the usefulness of this tool.

To remediate the shortcomings of chi-square based tests, a series of alternative fit indices (with corresponding cut-off criteria) has been developed. West, Taylor and Wu (2012[13]), for example, suggest relying on the root mean square error of approximation (RMSEA) and the comparative fit index (CFI). Simulations suggest that well-fitting models should provide RMSEA values which are smaller than 0.06 and CFI values which are higher than 0.95 (Hu and Bentler, 1999[14]). Yet, Chen (2007[15]) suggests that it is not sufficient that a model provides fit indices that fulfil these cut-off criteria. In addition, one needs to assess whether or not these global fit measures deteriorate to a large extent when moving from a configural to a metric invariant model and from a metric to a scalar invariant model. Since the chi-square difference test when moving from one level of invariance to the other may be too strict, especially when the sample size is large, Chen (2007[15]) suggests that the change in RMSEA should be smaller than 0.03, and the change in CFI should be smaller than 0.01 to be able to conclude that a higher level of measurement invariance is given.

A disadvantage of the MGCFA approach is that it is very strict, in the sense that it requires exact equality of parameters across groups (Davidov et al., 2015[16]). In real data analysis, exact equality of measurement parameters is almost never the case. When sample sizes are as large as they often are in cross-national survey research, substantively irrelevant measurement differences between groups lead to the conclusion that invariance cannot be established (Meuleman, 2012[17]; Oberski, 2014[18]). As a result, researchers are often confronted with the finding that scalar invariance is not supported by the data, and do not know whether they can rely on the estimated latent means. Different approaches have been suggested for how to deal with the problem of MGCFA being overly strict (Davidov et al., 2014[9]). For example, some researchers (Byrne, Shavelson and Muthén, 1989[6]; Steenkamp and Baumgartner, 1998[10]) suggest that measurement parameters do not need to be equal for all items, but that it is sufficient that only two items have equal parameters to be able to compare relationships and/or latent means. They call this situation "partial measurement invariance". Another approach is the recently developed alignment procedure which we explain in the next section.

The Alignment Procedure

The alignment procedure was developed by Asparouhov and Muthén (2014[19]). It allows estimating latent means even when exact equality of measurement parameters is not present in the data. Alignment begins at a similar starting point as the MGCFA approach: Observed indicators are seen as a linear function of a latent variable (with intercepts and factor loadings as measurement parameters; see Equation 2.1). The alignment procedure uses several steps. In a first step, the estimated model does not constrain factor loadings and intercepts to be equal across groups. As such, the alignment method relies on a configurally equivalent model where factor loadings and intercepts are allowed to differ across groups. Instead of constraining parameters across groups to be equal, however, a second step in the alignment procedure looks for a pattern of the parameter estimates in which differences between the measurement parameters are minimised [using a simplicity function; for more details see Asparouhov and Muthén (2014[19]). As a result, the procedure ends up with many minor differences between parameters and only a few large differences. Asparouhov and Muthén (2014[19]) indicate that this process is similar to a rotation in exploratory factor analysis. When the point is reached where the total amount of non-invariant parameters is minimised, the estimation stops and produces the measurement parameters including the latent means. These estimated latent means take the detected differences between factor loadings and intercepts across groups into account. Therefore, the estimated latent means provide the best possible comparability that can be achieved with the given data. The fit measures of the model are the same as in a configural invariance model.

Of course, the best possible comparability might still be insufficient to make valid comparisons possible. At this point, researchers may correctly ask whether one may rely on these estimated means. After all, according to the exact approach comparability requires completely equal factor loadings and intercepts. Asparouhov and Muthén provide a response to this crucial question. They conducted simulation studies and showed that if the share of parameters which are different across groups is 25% or lower, the means are probably comparable (Muthén and Asparouhov, 2014[20]). However, further simulations are needed to test the robustness of this assumption. Since the alignment output lists all the parameters that are unequal (a parameter is considered to be non-invariant if it differs significantly from the average of that parameter in the set of invariant groups), it is easy to count whether this number is higher or lower than 25% of the total number of factor loadings and intercept parameters.

The alignment procedure has further advantages besides estimating the most trustworthy means. First, it lists all the non-invariant parameters and researchers can easily identify them in the output. Indeed, some researchers may be interested to understand why they are not invariant. The fact that these parameters are clearly indicated by M*plus* (they are indicated between parentheses in the output) makes this job easy. Second, the alignment output lists the means and also provides a difference test for these means across the groups. In other words, it ranks the group means in a descending order and informs which ones are significantly different at 5%. Thus, it allows researchers to find out very quickly which groups score highest and which ones score lowest. It is a very convenient approach particularly (but not only) when there are many groups in the analysis.

Two further technical details are worth noting. First, alignment can be estimated in the frequentist approach using maximum likelihood or in the Bayesian approach. In the following example we will use maximum likelihood. The chapter about approximate invariance and Bayesian estimation (Chapter 4) will apply the Bayesian procedure for alignment. Second, the analysis with the alignment procedure allows using two estimation

options: Fixed and Free. In the free alignment all the latent means are freely estimated. In the fixed alignment the latent mean in one of the groups is fixed to zero. The free alignment may perform better (Asparouhov and Muthén, 2014[19]), but the authors admit that the model may not be identified. This was the case in our illustration. In this case, it is suggested to use the fixed option, as will be shown below. Muthén and Asparouhov (2018[21]), Cieciuch, Davidov and Schmidt (2018[22]) and Munck, Barber and Torney-Purta (2017[23]) provide further technical details and applications.

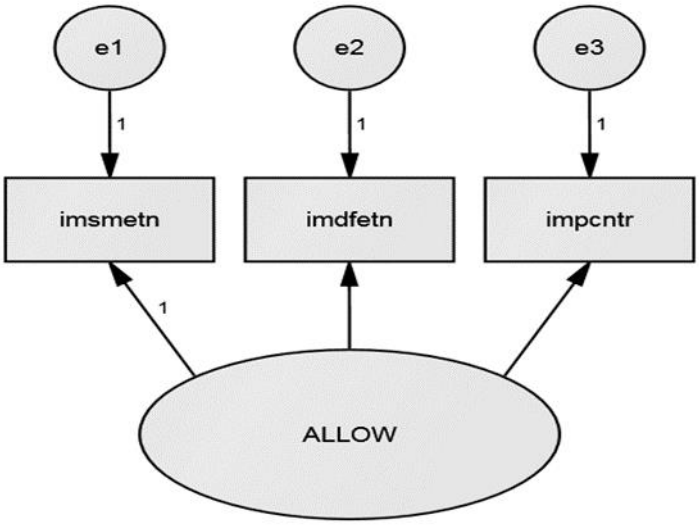
Illustration

Data and Measurements

To illustrate these procedures, we use ESS data collected in France covering all currently available seven rounds (2002, 2004, 2006, 2008, 2010, 2012 and 2014). The number of respondents in rounds 1-7 is 1 503, 1 806, 1 986, 2 073, 1 728, 1 968 and 1 917, respectively. Thus, the illustration presents and uses longitudinal (repeated cross-sectional) data in the country [for a similar approach, see Poznyak et al. (2013[24])]. Indeed, measurement invariance is important not only for the comparison of cultural groups, but also for the comparison of data collected at different time points for the same cultural group.

We used three items measuring the willingness of respondents to allow immigrants into the country (with a latent variable named ALLOW). The questions measuring this latent variable inquire whether respondents are willing to allow immigrants of the same race or ethnic group as the majority (*imsmetn*), of a different race or ethnic group than the majority (*imdfetn*), or from poorer countries outside Europe (*impcntr*) into the country. Responses ranged from 1 (allow many) to 4 (allow none). Thus, higher scores imply a stronger rejection of immigrants. Figure 2.2 presents the latent variable ALLOW and the three indicators measuring it.

Figure 2.2. A measurement model for the willingness to allow immigrants into the country



Results of the MGCFA Analysis

Table 2.1 presents the global fit indices for the MGCFA analysis and the three levels of exact measurement invariance that we tested: configural, metric and scalar. We used the software package Amos for the analysis. Looking at chi-square difference tests, moving from configural to metric and from metric to scalar invariance leads to a significant deterioration of model fit. As mentioned before, however, these chi-square based tests are very strict, and even substantively small deviations from invariance could lead to statistically significant misfit. According to the changes in alternative fit indices, such as CFI and RMSEA, moving from the configural to the metric invariance model, and from the metric to the scalar invariance model does not lead to a strong deterioration of model fit: the change in RMSEA is smaller than 0.03, and the change in CFI is smaller than 0.01 (see Table 2.1). As a result, one can conclude that measurement invariance is given on all levels based on the cut-off criterion suggested by Chen (2007[15]). Yet it has to be acknowledged that the evidence is not completely conclusive, as the chi-square difference tests point in the opposite direction.

Table 2.1. Model fit indices for the exact measurement invariance test using MGCFA

	Chi2	*df*	RMSEA	CFI
Configural	0	0		1.000
Metric	79.46	12	.021 [.017-.025]	.997
Scalar	232.59	24	.026 [.023-.029]	.990

Notes: Chi2 = chi-square; df = degrees of freedom; RMSEA = Root mean square error of approximation [with a 95% confidence interval]; CFI = Comparative fit index

Results of the Alignment Procedure

We begin the presentation with the syntax used to run the alignment procedure and explain each line of the syntax. For the analysis we used the software package M*plus* 7.4 (Muthén and Muthén, 1998-2017[25]). Table 2.2 presents the syntax and its explanation.

Table 2.2. The commands for running the alignment estimation procedure in M*plus*

Command	Explanation
data: file is FRANCE.dat;	*Defining the raw data file*
VARIABLE: names are essround imsmetn imdfetn impcntrb; usevariable imsmetn imdfetn impcntrb; missing = all (77 88 99);	*Lists the variables in the dataset and those used in the model. In addition, the missing values are defined.*
classes = c(7); knownclass = c(essround = 1 2 3 4 5 6 7);	*The alignment procedure uses a mixture model with a known number of classes (i.e., groups). The groups are ESS rounds in the current example.*
ANALYSIS: type = mixture;	*Defines a mixture analysis.*
Estimator = ML;	*Uses a maximum likelihood estimator.*
Alignment = fixed (6);	*Uses the fixed alignment.*
MODEL: %overall% ALLOW by imsmetn imdfetn impcntrb;	*Defines the model (a continuous latent variable measured by three indicators).*
OUTPUT: stand; tech1 tech8; align;	*Output request.*
SAVE: FILE IS align_FRANCE1_7.dat;	

Table 2.3 presents the non-invariant factor loadings and intercepts in the fixed alignment optimisation. These are listed in the M*plus* output. The list with non-invariant parameters indicates which parameters show substantially relevant deviations across groups, and is conceptually similar to the modification indices when running an MGCFA (and an exact measurement invariance test).

Table 2.3. Non-invariant parameters (factor loadings and intercepts)

	Factor Loadings			Intercepts		
ESS round	imsmetn	imdfetn	impcntrb	imsmetn	imdfetn	impcntrb
1						
2			X			
3						
4						
5						
6	X			X	X	X
7	X			X		X
Number of non-invariant parameters	2	0	1	2	1	2
percentage of non-invariant parameters	14%			24%		

Table 2.3 shows that 14% of the factor loading and 24% of the intercept parameters were non-invariant across groups. As a rule of thumb, Asparouhov and Muthén (2014[19]) put forward that meaningful comparisons could be made as long as the percentage of non-invariant parameters is lower than 25%. According to this heuristic, we can thus rely on the estimated means of the alignment procedure. Readers should be warned, however, that this is simply a rule of thumb based on a limited number of simulation studies. The percentage is a very rough indicator that does not take into account the pattern of differences, and is not sensitive to the size of the deviations from the average pattern. When strong deviations of equivalence are located in a limited number of groups, comparisons can be problematic

while the percentage of non-invariant parameters is misleadingly low. Therefore, a closer inspection of measurement parameters per group is advisable.

Next we present the means estimated in the alignment procedure. In Table 2.4 we present this part of the M*plus* output, and below Table 2.4 we interpret this information.

Table 2.4. M*plus* output comparing means of the latent variable ALLOW

Means of the latent variable ALLOW are presented in a descending order, along with statistically significant differences (at the 5% significance level)

Ranking	Group (ESS round)	Factor mean	Groups With Significantly Smaller Factor Mean
1	2	0.165	5 4 6 7
2	3	0.156	4 6 7
3	1	0.144	4 6 7
4	5	0.091	6 7
5	4	0.067	6 7
6	6	0	
7	7	-0.034	

An inspection of Table 2.4 reveals that groups 1, 2 and 3 (which correspond to ESS rounds 1, 2 and 3, i.e. years of data collection 2002, 2004 and 2006) have the highest means (i.e. the strongest rejection of immigrants). Their means are not significantly different from each other. The mean in round 2 (round 2004) is highest and is significantly higher than in ESS rounds 4, 5, 6 and 7. The means in ESS rounds 1 and 3 are significantly higher than the means in rounds 4, 6 and 7. The means in rounds 4 and 5 are not significantly different from each other and are significantly higher than those in rounds 6 and 7. The means in rounds 6 and 7 are the lowest and are not significantly different from each other. In other words, in these two last ESS rounds 6 and 7, rejection of immigration is lowest and people are the most willing to allow immigrants into the country.

Chapter 3. Bayesian Approximate Measurement Invariance

Kimberley Lek and Rens van de Schoot

Defaults versus Approximate Measurement Invariance

When measurement invariance (MI) does not hold, subjects from different groups (typically countries) or the same subjects at different time points respond differently to the items of a questionnaire. As a consequence, factor means cannot reasonably be compared, either across these groups or over time (Millsap, 2011[26]). Testing for MI is therefore a requirement when one wants to compare countries or time points on factor means. When there are many countries or time points involved, testing for MI is often a frustrating and cumbersome enterprise. Full MI rarely holds in such large datasets, and one is often confronted with many large, non-negligible modification indices. Moreover, problems may arise from the fact that releasing invariance constraints on the basis of these modification indices is not guaranteed to lead to the correct or simplest model, due to chance capitalisation (Muthén and Asparouhov, 2013[27]). So, what to do in such a situation?

Muthén and Asparouhov (2012[28]; 2013[29]) describe a novel method, labelled Bayesian structural equation modelling (BSEM), where exact zero constraints can be replaced with approximate zero constraints. BSEM is for instance used in the context of confirmatory factor analysis, where cross-loadings are traditionally constrained to be zero. By using the procedure of Muthén and Asparouhov (2012[28]), these cross-loadings can be estimated with some, as van de Schoot et al. (2013[30]) call it, 'wiggle room', implying that very small cross-loadings are allowed. Another area where approximate zeros might have an advantage is when full measurement invariance across groups does not hold, implying that exact zero differences between factor loadings and intercepts are too strict. Allowing small differences in factor loadings and/or intercepts can ensure a satisfactory model fit, termed Bayesian approximate MI. Bayesian approximate MI allows for some wiggle room for the intercept or factor loading differences between countries, where the wiggle room is determined by the degree of precision of the prior. The use of priors on the difference in parameters introduces a posterior distribution, which tries to find a compromise between the ideal situation (the difference between two parameters is zero) and the situation we find in the data (the difference is unrestricted). The willingness to compromise between model 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. For more details we refer to Van de Schoot et al. (2013[30]) and Lek et al. (2018[31]).

Bayesian approximate MI can be used in conjunction with the alignment method introduced in Chapter 2. The approximate MI solution is then rotated such that the number of non-invariant items is minimised. The choice for alignment depends on the anticipated structure of non-invariance in the data: approximate MI *without* alignment is suitable when there is a large degree of minor non-invariance, where the differences in intercepts and factor loadings largely cancel each other out (Asparouhov and Muthén, 2014[19]). Approximate MI *with* alignment is applicable when the majority of the items is invariant while a minority is not (Muthén and Asparouhov, 2013[27]). BSEM is currently available for situations where the latent variable is of continuous nature and when indicator variables are either continuous or could be approximated as continuous (e.g. Likert scales).

Illustration

Data and Measurements

To illustrate Bayesian approximate MI, we used ESS data collected in 22 countries (i.e. Austria, Belgium, Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Luxembourg, Netherlands, Norway, Poland, Portugal, Slovenia, Sweden and Switzerland) from the 2002 round. The total number of respondents in the 22 countries is 42 359 in this round with an average per country of 1 925 (min = 1 207; max = 2 919).

In accordance with Davidov and Meuleman [see Chapter 2; see also Meuleman & Billiet (2012[32]); Davidov et al. (2015[16])], we used three items measuring the willingness of respondents to allow immigrants into the country (with a latent variable named 'ALLOW'). The questions measuring this latent variable inquire whether respondents are willing to allow immigrants of the same race or ethnic group as the majority (*imsmetn*), from a different race or ethnic group than the majority (*imdfetn*), or from poorer countries outside Europe (*impcntr*) into the country. Responses ranged from 1 (allow many) to 4 (allow none). Thus, higher scores imply a stronger rejection of immigrants. Figure 2.2 in Chapter 2 presents the latent variable 'ALLOW' and the three indicators measuring it.

Analytic Strategy

The current illustration has two major goals. The first goal is to compare the Bayesian approximate MI solution to that of the traditional multigroup confirmatory factor approach (MGCFA) and (ML) alignment (see also Chapter 2), based on their factor mean ranking of the 22 countries. The second goal is to determine whether Bayesian approximate MI is feasible[1], with a prior on the *difference* in intercepts and slopes (e.g. factor loadings) across the 22 countries (for M*plus* code see Annex 3.A). The mean of this prior equals zero, because *on average* we want no differences in intercepts and slopes across groups. The variance of the prior determines the 'wiggle room' we allow in the intercept and factor loading estimates across the 22 countries. We use a prior variance of .05, but other values are possible and are compared in a sensitivity analysis (i.e. .001, .005, .01, .05, .1). In the absence of strict guidelines, we developed a simple procedure to investigate the Bayesian approximate MI solution, using the software R (R Core Team, 2017[33]) and the R package "MplusAutomation" (Hallquist and Wiley, 2018[34]); see Annex 3.B for the annotated R code.

Results

MGCFA

In Table 3.1 the results are displayed for the traditional configural, metric and scalar invariance models. Because of the large sample size, all Chi-square difference tests have p-values close to zero. Looking at the Chi-square values, the CFI and RMSEA [following the recommendation of Chen (2007[15])], the scalar invariance model seems problematic and full comparability of means is not achieved. According to the fit statistics, the scalar model is thus inappropriate to compare the latent means shown in the second column of

[1] Note we only report on the results of Bayesian approximate MI with alignment because approximate MI without alignment resulted, even after 100 000 iterations, in a posterior covariance matrix not positive definite in one of the chains.

Table 3.2 across the 22 countries. Note that these fit statistics may be too strict when the amount of non-invariance in the data is non-substantial.

Table 3.1. Model fit indices for the exact measurement invariance test using MGCFA

	Chi2	*df*	RMSEA	CFI
Configural	n/a	0	0	1.000
Metric	443.106	42	0.07 [0.065 0.076]	0.995
Scalar	1703.733	84	0.10 [0.096 0.104]	0.979

Note: Chi2 = chi-square; df = degrees of freedom; RMSEA = Root mean square error of approximation [with a 95% confidence interval]; CFI = Comparative fit index

Table 3.2. Comparison of the latent mean ordering across countries for MGCFA, BSEM with alignment and alignment with ML

	Factor mean (rank order)		
Country	MGCFA traditional	ML alignment	BSEM with alignment
Hungary	1.312 (1)	1.342 (1)	1.337 (1)
Greece	1.104 (2)	1.111 (2)	1.108 (2)
Luxembourg	1.005 (3)	1.028 (4)	1.024 (4)
Portugal	0.990 (4)	1.036 (3)	1.033 (3)
Spain	0.819 (5)	0.842 (7)	0.838 (7)
Israel	0.768 (6)	0.930 (6)	0.917 (6)
Austria	0.699 (7)	1.013 (5)	1.006 (5)
Czech Republic	0.676 (8)	0.711 (8)	0.707 (8)
Poland	0.554 (9)	0.565 (9)	0.563 (9)
Denmark	0.511 (10)	0.525 (10)	0.521 (10)
France	0.477 (11)	0.518 (11)	0.514 (11)
Finland	0.476 (12)	0.485 (12)	0.482 (12)
United Kingdom	0.366 (13)	0.384 (14)	0.382 (14)
Slovenia	0.347 (14)	0.482 (13)	0.479 (13)
Belgium	0.349 (15)	0.358 (15)	0.356 (15)
Italy	0.296 (16)	0.312 (16)	0.310 (16)
Netherlands	0.278 (17)	0.307 (17)	0.305 (17)
Germany	0.180 (18)	0.182 (19)	0.182 (19)
Ireland	0.177 (19)	0.184 (18)	0.179 (18)
Switzerland	0.124 (20)	0.128 (20)	0.126 (20)
Norway	0.000 (21)	0.000 (21)	0.000 (21)
Sweden	-0.222 (22)	-0.224 (22)	-0.225 (22)

Notes: MGCFA traditional wrongly assumes scalar measurement invariance. Norway is used as the reference group with factor mean 0 (and factor variance 1).

Alignment (ML)

The third column of Table 3.2 contains the estimated factor means and their ranking when using the ML alignment method (see previous chapter for more details). In order to compare these factor means, the model should have a satisfactory model fit *and* the majority of the items should be non-invariant. With regard to model fit, we have zero degrees of freedom to obtain model fit indices (due to the small number of items). We therefore simply assume – for this illustration – that our alignment ML model fits the data. With regard to the degree of (non)invariance, Muthén and Asparouhov (2014[20]) suggest 25% of the parameters to be non-invariant or less as a general rule of thumb. The M*plus* output labelled

"Approximate measurement invariance (non-invariance) for groups" indicates that in our illustration, 21 intercepts and 16 factor loadings appear to be non-invariant over the 22 countries, leading to an average of 28% non-invariance (31.82% factor loading non-invariance; 24.24% intercept non-invariance).

Bayesian Approximate MI with Alignment

The fourth column of Table 3.2 shows the result of approximate MI with the alignment option. Again, before comparing the factor means, model fit should be satisfactory and the majority of the items should be invariant. To assess model fit, we relied on the posterior predictive p-value (PPP). PPP-values around 0.5 indicate a good predictive model fit. The PPP in our illustration is 0.503. It is tempting to evaluate small variance priors using readily available approaches like the posterior predictive p-value and the Deviance Information Criterion (DIC). However, as was shown by Hoijtink and Van de Schoot (2018[35]) both are not really suited for the evaluation of models based on small variance priors. An alternative is the prior-posterior predictive p-value, which is currently being implemented in software.

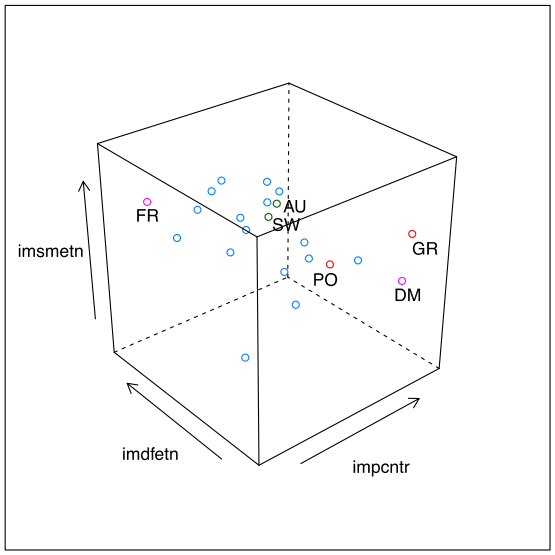
Prior choice

Ideally, the estimated differences in latent means over the 22 countries should not depend on the chosen prior variance. To investigate the influence of prior variance choice, we look at the latent mean difference estimates we would obtain with different reasonable prior variances. When the number of countries is large, as in our example, it can be infeasible to check this prior influence for every combination of countries. Therefore, we limit ourselves to three comparisons: Switzerland (SW) versus Austria (AU), Greece (GR) versus Portugal (PO) and Denmark (DM) versus France (FR). We choose these three comparisons based on Figure 3.1, which plots the Euclidean distance for the countries' intercept*factor loading values for each of the three items at a prior variance .05 (the annotated R code in Annex 3.B enables us to compute this Euclidean distance for each country indicating the level of non-invariance when compared to other countries.).

In Figure 3.1, SW and AU respectively show the smallest Euclidean distance indicating these two countries have very similar intercepts and factor loadings and hence are rather similar in their level of non-invariance. PO and GR show a median Euclidean distance and DM and FR show the largest Euclidean distance indicating these two countries are least similar in the combination of the estimates for the intercepts and factor loadings. Checking this last combination with the largest distance is particularly important (see below).

Figure 3.1. Visual representation of the variability in model parameters in an approximate invariance model

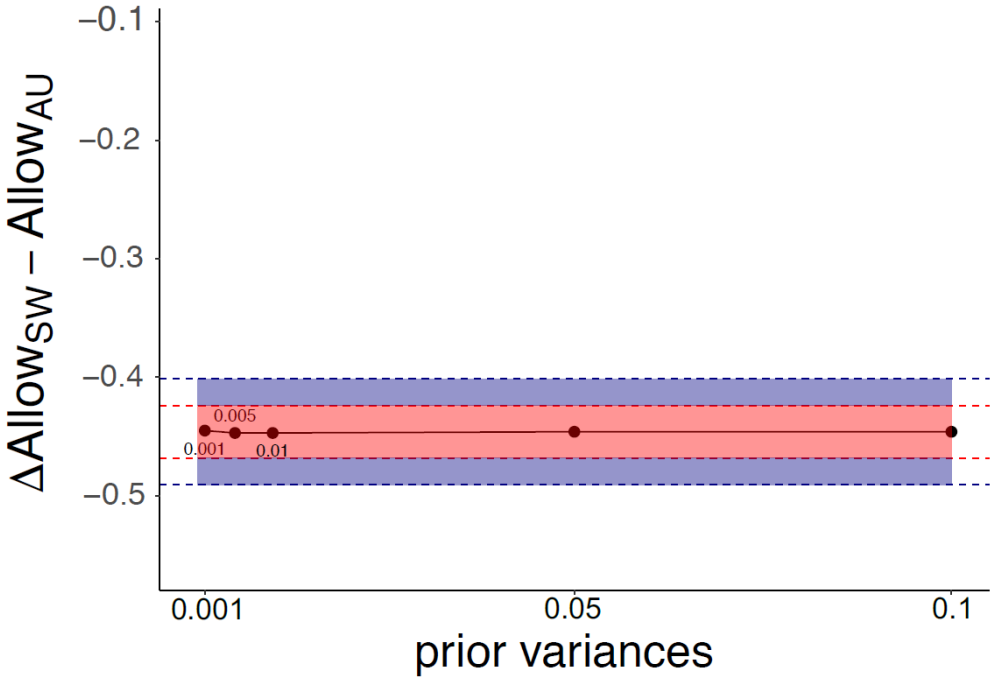
intercept * factor loading values are plotted for each of the three items as estimated with a prior variance of .05



Notes: See Annex 3.B for computational details. The green dots show the smallest Euclidian distance between countries (AU and SW), the red dots the median Euclidean distance (PO and GR) and the pink dots the largest Euclidean distance (FR and DM).

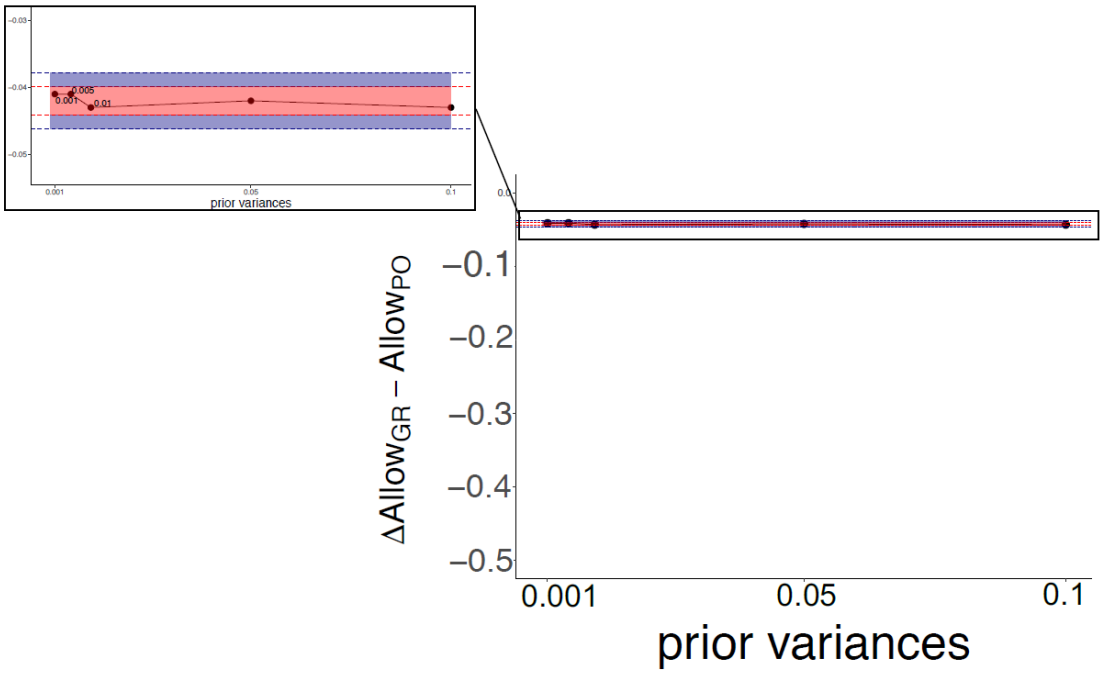
For each of the three country combinations, Figure 3.2, Figure 3.3 and Figure 3.4 illustrate the influence of prior variance choice on the estimated latent mean difference. When factor loadings and intercept differences between countries are relatively small (SW versus AU; Figure 3.2) or average (GR versus PO; Figure 3.3), all estimated differences fall within a small range. This can be used as an indication that the difference in ALLOW estimates is reasonably robust against changes in prior variance for countries that do not differ too much in terms of factor loadings and intercepts. When factor loadings and intercepts differ considerably, as with Denmark and France (Figure 3.4), the estimates differ, particularly for relatively small prior variances. Based on this information, one would want to understand why France and Denmark have such a different interpretation of the items.

Figure 3.2. Differences in the means of "ALLOW" for Switzerland and Austria, by prior variance



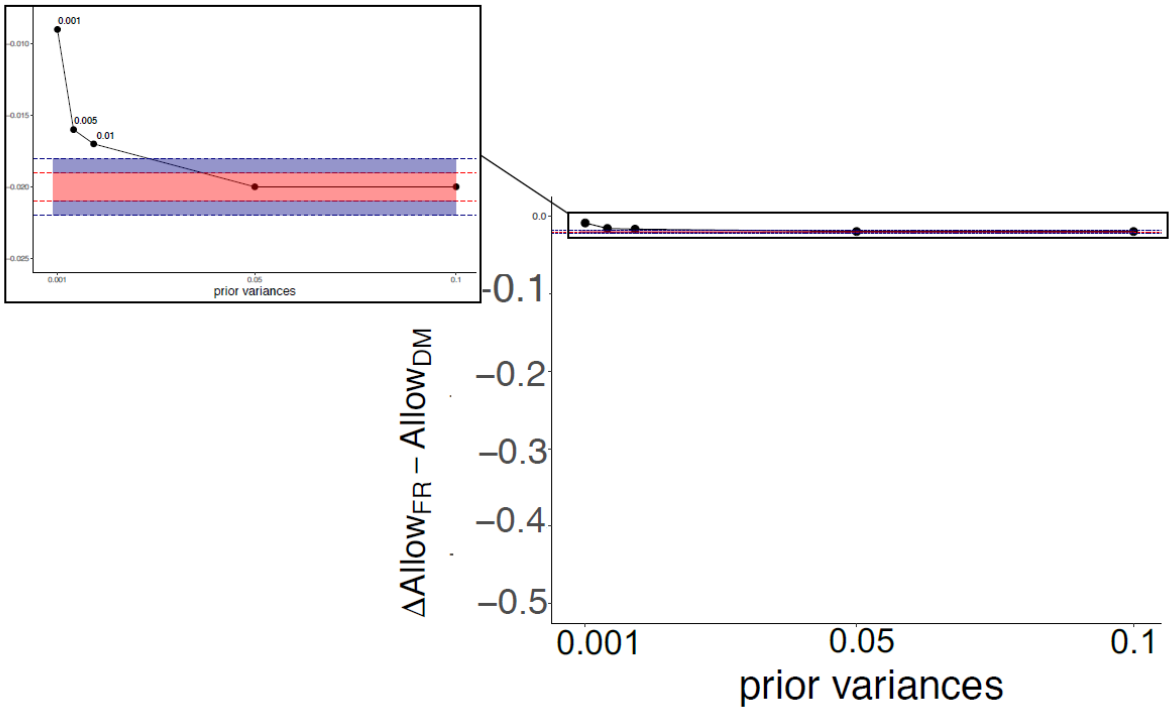
Notes: Estimates correspond to prior variances .001, .005, .01, .05 and .1. Estimates are connected by line segments to ease interpretation. The y-axis range is -0.5 to 0.

Figure 3.3. Differences in the means of "ALLOW" for Greece and Poland, by prior variance



Notes: Estimates correspond to prior variances .001, .005, .01, .05 and .1. Estimates are connected by line segments to ease interpretation. The y-axis range is -0.5 to 0 in the right panel, and -0.05 to -0.03 in the left panel.

Figure 3.4. Differences in the means of "ALLOW" for France and Denmark, by prior variance



Notes: Estimates correspond to prior variances .001, .005, .01, .05 and .1. Estimates are connected by line segments to ease interpretation. The y-axis range is -0.5 to 0 in the right panel, and -0.025 to -0.01 in the left panel.

Discussion

When comparing latent factor means across many countries, often the test for measurement invariance fails, as was the case in our illustration: the scalar model did not reach satisfactory model fit. As was argued in the previous chapter, a solution might be to use the alignment method. As we demonstrated, the results of the alignment method can still result in too many of the item-country combinations being non-invariant. A solution suggested in the literature is the method of approximate measurement invariance which reduces the level of non-invariance by using the Bayesian toolbox. Using strict priors on the parameter differences between countries, the non-invariance is "gently" pushed towards zero leaving some wiggle room and thereby avoiding exact invariance.

Although the factor means between the three methods are quite comparable (with correlations between the factor means larger than .98), there are a few ranking differences (see Table 3.2). Most of the changes in rank order that occurred between the first model in comparison with the other two happened in only few countries: Austria, Israel and Slovenia. Apparently, the scores for these countries are most influenced by the choice of statistical model. However, the highest degree of non-invariance in parameters is found when comparing France and Denmark. Based on the results of the three different

approaches, only the factor means and rank order of the Bayesian approximate model with alignment should be used for interpretation and further analyses with the exception of the comparison of France and Denmark. The next step is to come up with explanations why the participants in France and Denmark interpreted the questions in a different way. Finding a good explanation would require further study. The different interpretations are unlikely to be due to translation problems, given the rigorous translation procedures used in the ESS project.

Recommendations

When there are many countries or time points in our data, full measurement invariance rarely holds. Bayesian approximate MI with alignment can be a viable alternative in these instances, balancing theory (no differences in factor loadings and intercepts across countries or time points) and reality (model fit). As the method of Bayesian approximate MI is relatively new, there are no strict guidelines yet to determine whether approximate MI holds or which prior settings to use. Therefore, we advise performing a sensitivity analysis for country combinations with the largest non-invariance as is estimated with the Euclidian Distance (see Annex 3.B). Model fit indices to assist making decisions on model fit with informative priors and which prior settings to use are currently being developed and slowly being integrated in software (Hoijtink and van de Schoot, 2018[35]). All in all, the method of approximate MI is promising, especially in the case with many small deviating items, but the application to empirical data should be carefully done.

Annex 3.A. M*plus* Input File

Annex Table 3.A.1. Input file Bayesian approximate MI with alignment

Syntax	Explanation
DATA: FILE = "ESSdata.dat" ;	*Defining the data file with 22 countries and 1 wave (2002).*
VARIABLE:	
NAMES ARE imsmetn imdfetn impcntr imbgeco imueclt imwbcnt newctry;	*Variable names in the dataset,*
USEVARIABLES ARE imsmetn imdfetn impcntr;	*variables used for the MGCFA*
MISSING = all (77 88 99);	*missing data specification*
classes = g(22);	*The 22 countries are defined as known classes, in a mixture analysis (newctry contains the numbers for the 22 countries)*
KNOWNCLASS IS g (newctry = 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22);	
ANALYSIS:	*For the alignment procedure, a mixture model is specified. Here, we use fixed alignment, where the factor mean (and factor variance) of the 18th country (Norway) are constrained.*
type is mixture;	
alignment = fixed (18 BSEM);	
estimator is bayes;	*For this illustration, Bayesian statistics is used together with the alignment statement. The Bconvergence, Biterations and chains options all aid convergence.*
Bconvergence=0.05;	
Biterations = 500000(100000);	
processor = 2;	
chains = 2;	
BSEED = 123;	
MODEL:	
%OVERALL%	*overall model statement*
Allow by imsmetn;	
Allow by imdfetn;	
Allow by impcntr;	
[imsmetn imdfetn impcntr];	
%G#1%	*This part is repeated for every of the G countries (here illustrated for country 1 = Austria). Note the labeling of the factor loadings (first number is for the group, second number for the item) and the intercepts.*
Allow by imsmetn (lam1_1);	
Allow by imdfetn (lam1_2);	
Allow by impcntr (lam1_3);	
[imsmetn] (nu1_1);	
[imdfetn] (nu1_2);	
[impcntr] (nu1_3);	
[…]	
MODEL PRIORS:	*In this part, priors are placed on the differences in intercepts and factor loadings across the countries. DO(1,3) implies that the placement of the prior should be done for item 1,2 and 3. DIFF makes sure the prior is placed on the difference in factor loadings and intercepts, not on the factor loadings and intercepts themselves. # is being replaced by 1,2 and 3 in the DO loop.*
DO (1,3) DIFF(lam1_#-lam22_#) ~ N(0,0.05);	
DO (1,3) DIFF(nu1_#-nu22_#) ~ N(0,0.05);	
OUTPUT: Align;	

Annex 3.B. R code

```
############### Packages ######################

# install.packages("MplusAutomation")
require(MplusAutomation)

# install.packages("ggplot2")
require(ggplot2)

# install.packages("lattice")
require(lattice)

###### step 1.] create Mplus input files #####

## templateFile is a .txt file based on 'template language' (see vignette "MplusAutomation")
## to automatically create input files for every possible prior variance

createModels("F:/MplusAutomationOutput/templateFile.txt")

###### step 2.] run models from input files #####

runModels("F:/MplusAutomationOutput")
```

```
###### step 3.] extract model parameters ###########

variances <- c(0.001, 0.005, 0.01, 0.05, 0.1)     # prior variances


## Storage ##
factor.l.1 <- list()    # factor loadings item 1 ("IMSMETN")
interc.1   <- list()    # intercepts item 1 ("IMSMETN")

factor.l.2 <- list()    # factor loadings item 2 ("IMDFETN")
interc.2   <- list()    # itercepts item 2 ("IMDFETN")

factor.l.3 <- list()    # factor loadings item 3 ("IMPCNTR")
interc.3   <- list()    # intercepts item 3 ("IMPCNTR")

   f.means <- list()    # factor means for ALLOW



model.estimation <- numeric(length(variances))

for (i in 1:length(variances)) {

  ## Extract all parameters from created output files
              path <- paste("/Users/Kimberley/Desktop/MplusAutomationOutput/",variances[i]," -
                            approximate mi input.out", sep = "")
  Extract.Parameters <- extractModelParameters(path)$unstandardized
          parameters <- cbind(Extract.Parameters[["paramHeader"]], Extract.Parameters[["param"]],
                            Extract.Parameters[["est"]],Extract.Parameters[["posterior_sd"]])

  ## item IMSMETN
  factor.l.1[[i]] <- parameters[parameters[,1] == "ALLOW.BY"   & parameters[,2] == "IMSMETN", ]
    interc.1[[i]] <- parameters[parameters[,1] == "Intercepts" & parameters[,2] == "IMSMETN", ]

  ## item IMDFETN
  factor.l.2[[i]] <- parameters[parameters[,1] == "ALLOW.BY"   & parameters[,2] == "IMDFETN", ]
    interc.2[[i]] <- parameters[parameters[,1] == "Intercepts" & parameters[,2] == "IMDFETN", ]

  ## item IMPCNTR
  factor.l.3[[i]] <- parameters[parameters[,1] == "ALLOW.BY"   & parameters[,2] == "IMPCNTR", ]
    interc.3[[i]] <- parameters[parameters[,1] == "Intercepts" & parameters[,2] == "IMPCNTR", ]

     f.means[[i]] <- parameters[parameters[,1] == "Means" & parameters[,2] == "ALLOW", ]


  ### check minimal requirement of convergence ###
  if (ncol(factor.l.3[[i]]) == 4) {model.estimation[i] <- 0}  # model estimation terminated normally
  if (ncol(factor.l.3[[i]]) == 3) {model.estimation[i] <- 1}  # model estimation did NOT terminate normally (no
fourth column with posterior sds)
}
```

```
################ step 4.] Determine Euclidean Distance ###################


# variance = 4 is 0.05; the reference variance

#### factor loadings * intercepts #####

## item IMSMETN
impact.1 <- as.numeric(factor.l.1[[4]][1:22,3]) * as.numeric(interc.1[[4]][1:22,3])

## item IMDFETN
impact.2 <- as.numeric(factor.l.2[[4]][1:22,3]) * as.numeric(interc.2[[4]][1:22,3])

## item IMPCNTR
impact.3 <- as.numeric(factor.l.3[[4]][1:22,3]) * as.numeric(interc.3[[4]][1:22,3])

## determine Euclidean distance
X <- cbind(impact.1, impact.2, impact.3)
Euclidean.distances <- as.matrix(dist(X, method = "euclidean"))

# the diagonal is set to NA because we are not interested in the distances from a country "to itself"
diag(Euclidean.distances) <- NA

for (r in 1:22) {          # rows
  for (c in 1:22) {        # columns
    if (r != c) {          # not interested in diagonal

      # row & column of countries with largest Euclidean distance
      if (Euclidean.distances[r,c] == max(Euclidean.distances, na.rm = TRUE)){nrs <- c(r, c)}

      # row & column of countries with smallest Euclidean distance
      if (Euclidean.distances[r,c] == min(Euclidean.distances, na.rm = TRUE)){smallest <- c(r, c)}

      # row & column of countries with median Euclidean distance
      if (Euclidean.distances[r,c] == median(Euclidean.distances, na.rm = TRUE)){middle <- c(r, c)}

      }
  }
}
```

```
############################### Euclidean distances plot ###################################

imsmetn <- impact.1
imdfetn <- impact.2
impcntr <- impact.3

# Different colors for smallest, median and largest Euclidean Distance countries
COL <- numeric(22)

for (i in 1:22) {
  if (i ==      nrs[1] | i ==      nrs[2]) {COL[i] <- 1}
  if (i == smallest[1] | i == smallest[2]) {COL[i] <- 2}
  if (i ==   middle[1] | i ==   middle[2]) {COL[i] <- 3}
}


datA <- data.frame(cbind(imsmetn,imdfetn, impcntr, COL))

wireframe(imsmetn ~ impcntr * imdfetn, data=datA)
cloud(imsmetn ~ impcntr * imdfetn, data=datA, group = COL, auto.key = F, main = "intercept * factorloading")

# Optional: manually add names of countries
grid::grid.text("FR", x=unit(0.35, "npc"), y=unit(0.59, "npc"))
grid::grid.text("DM", x=unit(0.65, "npc"), y=unit(0.46, "npc"))
grid::grid.text("AU", x=unit(0.52, "npc"), y=unit(0.605, "npc"))
grid::grid.text("SW", x=unit(0.51, "npc"), y=unit(0.575, "npc"))
grid::grid.text("PO", x=unit(0.555, "npc"), y=unit(0.485, "npc"))
grid::grid.text("GR", x=unit(0.67, "npc"), y=unit(0.536, "npc"))
```

```
############ Figure 2 #############

# A: latent mean differences largest

# note: we only show the code for Figure 2A here, but Figure 2B and 2C can easily be found
# by changing "nrs[1]" and "nrs[2]" (the numbers of the countries with the largest Euclidean D.)
# to "smallest[1] & smallest[2]" and "middle[1] & middle[2]".

## benchmark difference in ALLOW with prior variance .05
benchmark <- as.numeric(f.means[[4]][nrs[1],3])-as.numeric(f.means[[4]][nrs[2],3])

## mean differences for other values of prior variance
diff <- numeric(5)

for (d in 1:5){ diff[d] <- as.numeric(f.means[[d]][nrs[1],3])-as.numeric(f.means[[d]][nrs[2],3]) }

# Red and blue area
five.procent <- benchmark / 100 * 5
 ten.procent <- benchmark / 100 * 10

# used for scale of the plot (y-axis runs from benchmark - 25% to benchmark + 75%)
 twentyfive.procent <- benchmark / 100 * 25
seventyfive.procent <- benchmark / 100 * 75

y <- diff
x <- variances
df <- as.data.frame(cbind(x,y))
```

```
PLOT1 <- ggplot(data = df, aes(x = df$x, y = df$y)) + geom_point(size = 3, type = 4) + geom_line() + theme_class
ic() + theme(
  axis.line.x = element_line(color="black", size = 0.5),
  axis.line.y = element_line(color="black", size = 0.5),
  panel.grid.major = element_blank(),
  panel.grid.minor = element_blank(),
  panel.border = element_blank(),
  panel.background = element_blank(),
  axis.title.y = element_text(size=15),
  axis.title.x = element_text(size=15)) +
  scale_x_continuous(limits = c(0,0.1), breaks = c(0.001, 0.05, 0.1), labels = c(as.character(c(0.001, 0.05, 0.1
)))) + ylab(expression(paste(Delta, "Allow"["FR"] - "Allow"["DM"]))) + xlab("prior variances") +
  geom_hline(yintercept = benchmark + five.procent, color = "red", linetype = 2) +
  geom_hline(yintercept = benchmark - five.procent, color = "red", linetype = 2) +
  geom_hline(yintercept = benchmark + ten.procent, color = "navyblue", linetype = 2) +
  geom_hline(yintercept = benchmark - ten.procent, color = "navyblue", linetype = 2)  +
  geom_rect(aes(xmin=0, xmax=0.1, ymin=benchmark - five.procent, ymax=benchmark + five.procent), fill="red", alp
ha=0.1, inherit.aes = FALSE) +
  geom_rect(aes(xmin=0, xmax=0.1, ymin=benchmark - ten.procent, ymax=benchmark - five.procent), fill="navyblue",
alpha=0.1, inherit.aes = FALSE) +
  geom_rect(aes(xmin=0, xmax=0.1, ymin=benchmark + five.procent, ymax=benchmark + ten.procent), fill="navyblue",
alpha=0.1, inherit.aes = FALSE) +
  annotate(geom = "text", x = variances[1]+0.003, y = y[1]+.0006, color = "black", label = as.character(x[1]), s
ize = 4) +
  annotate(geom = "text", x = variances[2]+0.003, y = y[2]+.0006, color = "black", label = as.character(x[2]), s
ize = 4) +
  annotate(geom = "text", x = variances[3]+0.003, y = y[3]+.0006, color = "black", label = as.character(x[3]), s
ize = 4) + ylim(c(benchmark + twentyfive.procent, benchmark - seventyfive.procent))
```

templateFile.txt

```
[[init]]
iterators = variances;
variances = 0.001 0.005 0.01 0.05 0.1;
filename = "[[variances]] - Approximate MI input.inp";
outputDirectory = "F:/MplusAutomationOutput";
[[/init]]
```

```
title: Application approximate MI

 DATA:  FILE = "F:/MplusAutomationOutput/ESSdata.dat" ;
 VARIABLE:  NAMES ARE imsmetn imdfetn impcntr imbgeco imueclt imwbcnt newctry;
  usevariables are imsmetn imdfetn impcntr;
  missing = all (77 88 99);
  classes = g(22);
  KNOWNCLASS IS g (newctry = 1 2 3 4 5 6 7 8 9 10 11 12
                                      13 14 15 16 17 18 19 20 21 22);

    ANALYSIS:
    !model is allfree;
    alignment = fixed (bsem);
    type is mixture;
    estimator is bayes;
    Bconvergence=0.01;
    Biterations = 500000(100000);
    processor = 2;
    chains = 2;
    bseed = 123;

    MODEL:

      %OVERALL%
      Allow by imsmetn;
      Allow by imdfetn;
      Allow by impcntr;

  [imsmetn imdfetn impcntr];

      %G#1%
      Allow by imsmetn (lam1_1);
      Allow by imdfetn (lam1_2);
      Allow by impcntr (lam1_3);

      [imsmetn] (nu1_1);
      [imdfetn] (nu1_2);
      [impcntr] (nu1_3);

!!!!!!!!!! repeated for %G#2% - %G#22% !!!!!!!!!!


    MODEL PRIORS:
  DO (1,3) DIFF(lam1_#-lam22_#) ~ N(0,[[variances]]);
  DO (1,3) DIFF(nu1_#-nu22_#) ~ N(0,[[variances]]);


    PLOT:
    type is plot2;

  OUTPUT: ALIGN;
```

Chapter 4. Cross-Cultural Comparability in Questionnaire Scales: Bayesian Marginal Measurement Invariance Testing

Jean-Paul Fox

Introduction

When administering a test to different groups, it is important to be able to compare the test results across members of those groups. In order to make meaningful comparisons between groups, the latent variable θ (i.e. ability or propensity) must be measured on a common scale. To accomplish a common scale analysis, the possible violation of the assumption of measurement invariance should be taken into account (Thissen, Steinberg and Gerrard, 1986[36]; Fox, 2010[37]; van de Vijver and Tanzer, 2004[38]). In item response theory (IRT), measurement invariance is present when the conditional probability of answering an item correctly does not depend on group information (Thissen, Steinberg and Gerrard, 1986[36]).

In current Bayesian methods, random item effects are used to detect measurement non-invariance. More specifically, deviations from the overall mean are specified for each group-specific item parameter (Fox, 2010, pp. 193-225[37]; Kelcey, McGinn and Hill, 2014[39]). The variance between groups with respect to these deviations is evaluated in order to detect measurement non-invariance: The larger the variance between groups, the higher the degree of measurement variance. These current methods are based on a conditional IRT modelling approach, where inferences are made regarding the latent variable conditional on the estimates of the group-specific item parameters. Verhagen and Fox (2013[40]) showed that Bayesian methods can be used concurrently to test multiple invariance hypotheses for groups randomly sampled from a population. They found that a Bayes factor test had good power and low Type I error rates for different sample size conditions to detect measurement non-invariance.

The random item effects approach is not suitable, when only a few groups are considered which are not sampled from a larger population. For a few groups, the between-group variance cannot be accurately estimated. Furthermore, this variance parameter has no correct interpretation when the selected groups define the entire population. In practice, in the sample design, groups are often considered to be fixed units (i.e. strata), and there is a specific interest in the selected groups, which constitute the entire population. A well-known two-group setting is the comparison of a single focal group to a single reference group, where a grouping variable (e.g. gender or geographic location) is the subgroup-classification or stratification variable. For non-randomly sampled groups (strata), Verhagen et al. (2016[41]) proposed another Bayes factor test, which was able to directly evaluate item difficulty parameter differences across the selected groups.

The current Bayesian approaches have several limitations. First, the variance between group-specific item parameters is explicitly modelled even though the object of these methods is to test whether this variance is present, which would indicate that the measurement invariance assumption is violated. That is, the prior for the variance parameter reflects an assumption of measurement non-invariance. Second, the model representing measurement invariance is not nested within the model representing measurement non-invariance. Measurement invariance is represented by a variance of zero, which is a boundary value on the parameter space (Fox, Mulder and Sinharay,

2017[42]).This complicates statistical test procedures and requires approximate methods such as an encompassing prior approach (Klugkist and Hoijtink, 2007[43]). Third, the latent variable θ is estimated using potentially biased item difficulty and population parameter estimates. Fourth, the above-mentioned approaches are applicable either to a non-randomly selected number of groups (strata) or to randomly selected groups (clusters), but none of the approaches is applicable to both situations.

Van de Schoot et al. (2013[30]) introduced a different Bayesian approach, where a prior distribution is specified for the 'invariant' item parameters, allowing them to vary across groups. The prior distribution for the item parameter provides support to variability in item parameter values across groups. When the prior variance is sufficiently small, approximate measurement invariance is considered. They also demonstrated that this prior for the item parameters can be used to evaluate approximate measurement invariance and to determine acceptable differences in item functioning between groups.

This method also has several drawbacks. First, the variance of the prior distribution determines the level of possible variation in item functioning, which needs to be specified a priori. In general, the magnitude of non-invariance for each item is usually unknown. Second, the prior is cantered around zero, where the point zero represents measurement invariance. The shape of the prior distribution can easily favour the measurement invariance assumption over the non-invariance assumption. For instance, when the prior distribution is single-peaked and cantered around the mean value (e.g. a normal distribution with mean zero), the point zero, representing measurement invariance, is a priori favoured over any other point representing non-invariance. Third, models with different prior variances for the item parameters do not differ in their number of model parameters, which complicates the model selection procedure. For instance, the less restrictive model with larger prior variances, but an equal number of model parameters as the one with smaller prior variances, will always be favoured by the usual information criteria such as the Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) (Kim et al., 2017[44]; van de Schoot et al., 2013[30]). Fourth, the specified prior variance to represent approximate measurement invariance depends on the sample size. In Kim et al. (2017[44]) and van de Schoot et al. (2013[30]), a prior variance of .001 represents approximate measurement invariance. Davidov et al. (2015[16]) allowed a variance of .05 under the approximate measurement invariance assumption. Specifically, for smaller sample sizes, the approximate measurement invariance model is often selected over the true model with a prior variance of .05. As the sample size decreases, the prior variance of .001 will lead to more shrinkage of the posterior mean estimate towards the prior mean, representing approximate measurement invariance. Therefore, when sample sizes are small, the prior variance representing approximate measurement invariance, can easily represent overwhelming evidence in favour of the measurement invariance hypothesis. In the same way, it is not possible to identify a specific prior variance as the allowed magnitude of variation that is acceptable as approximate measurement invariance, since the influence of the prior variance is sample dependent.

To overcome these limitations, a new method based on a marginal random item effects model is proposed. Instead of conditioning on group-specific item parameters, common item parameters (i.e. a measurement invariant item parameters) are modelled, which apply to all groups. As a result, the possible error with respect to these item parameters is included in the residuals for each group. It is proposed that in order to detect measurement non-invariance, the correlation of within-group residuals should be evaluated. Hence, the additional correlation between observations caused by violations of measurement invariance is addressed in the marginal random item effects model. Additionally, since

residual correlations between response probabilities are evaluated, the complex identification assumptions associated with the random item effects model (De Jong, Steenkamp and Fox, 2007[45]; Verhagen and Fox, 2013[40]) can be avoided. A further benefit of the proposed method is that it can be applied to both randomly selected groups (clusters) and non-randomly selected groups (strata) to make inferences about measurement invariance in the population and for the groups in the sample, respectively.

The current description of the marginal model is based on a random item difficulty parameter as described in Fox, Mulder and Sinharay (2017[42]). Before discussing in detail the marginal test approach, a comparison is made with traditional tests for measurement invariance and score purification methods. Then, the performance of the method is shown through simulation studies and an illustration of the method is given using a real data example. Finally, a discussion is given to show possible extensions of the method to more general situations. This includes a marginal model for the two-parameter IRT model to show that the measurement invariance test approach can be extended to more advanced IRT models.

Differential Item Functioning Methods

One class of methods for the detection of measurement non-invariance, or differential item functioning (DIF), can be considered as methods based directly on observable statistics. Within this class, the Mantel-Haenszel (MH) procedure is one of the most common procedures used to identify DIF (Holland and Wainer, 1993[4]). It can be computed to detect measurement non-invariance between two groups, where one group is called the *reference group* and the other group is called the *focal group*. The reference and the focal group are compared in terms of a matching variable (e.g. total test score or subscore) so that comparable groups are formed. Score equivalence between groups is established through a matching variable.

Under right-scoring for an item, a contingency table can be created for each subgroup (Hambleton and Rogers, 1989[46]).

Figure 4.1. Contingency table for a binary item, with groups matched on g

	Incorrect (0)	Correct (1)	
Reference Group	A_g	B_g	A_g+B_g
Focal Group	C_g	D_g	C_g+D_g
	A_g+C_g	B_g+D_g	N_g

The MH statistic evaluates whether the odds of getting the item correct at a given level of the matching variable is the same in the focal and reference groups.

A violation of measurement invariance is represented by an item-by-subgroup interaction. When the item is measurement invariant, the distribution of the item responses does not depend on group membership for each level of the matching variable. As a consequence, conditional on each level of the total score, responses to the measurement invariant item

are assumed to be a simple random sample. Subsequently, a violation of measurement invariance corresponds to a violation of the basic assumption of independence of a simple random sample. The MH method evaluates whether there is an interaction between the group and the item responses, which corresponds to evaluating the independence assumption of a simple random sample. If this assumption is violated, a cluster (or stratified) sample is observed instead of a simple random sample, and measurement invariance does not hold.

The MH statistic is computed as follows (Hambleton and Rogers, 1989[46]; Magis, Béland and Raiche, 2015[47]):

$$\chi^2_{\mathrm{MH}} = \frac{\left(\left| \sum_{g=1}^{G} A_g - \sum_{g=1}^{G} E(A_g) \right| - \frac{1}{2} \right)^2}{\sum_{g=1}^{G} V(A_g)}, \qquad \textbf{Equation 4.1}$$

where

$$E\left(A_g\right) = \frac{\left(A_g + C_g\right) \cdot \left(A_g + B_g\right)}{N_g} \qquad \textbf{Equation 4.2}$$

and

$$V\left(A_g\right) = \frac{\left(A_g + C_g\right) \cdot \left(B_g + D_g\right) \cdot \left(A_g + B_g\right) \cdot \left(C_g + D_g\right)}{N_g^2 \cdot \left(N_g - 1\right)}. \qquad \textbf{Equation 4.3}$$

The χ^2_{MH} statistic is assumed to be chi-squared distributed under the assumption of measurement invariance (i.e. absence of DIF). A violation of measurement invariance is detected, when the test statistic is considered extreme under the chi-squared distribution.

The MH test has several disadvantages. First, it is based on a two-group comparison, which makes it impossible to test simultaneously measurement invariance across multiple groups. In a multigroup situation, all subsets of two groups need to be compared. This will provide information about the violation of measurement invariance for each subset of two groups. The MH test results can contradict each other, where for instance a violation is detected for group 1 and 2, but not for group 1 and 3 and group 2 and 3.

Second, the MH statistic cannot handle randomly selected groups from a larger population, and will only provide information about a violation of measurement invariance related to the two groups. For randomly selected groups, the object of interest is whether the item can be marked as measurement invariant for the considered population. For instance, assume an item shows a violation of measurement invariance for a sample of randomly selected countries from the European Union. Then, it would be of interest to know that the item can also be marked as a DIF item for the other countries in the European Union that were not selected in the sample. However, this is not possible with the MH test, since the MH test results cannot be generalised to a larger population from which the sample was taken.

Third, the extremeness of the MH statistic is based on a chi-squared sampling distribution, which is an asymptotic approximation. The approximation might be poor, when only a few levels of the matching variable are considered (i.e. a few test items) and/or when sample sizes are small.

Fourth, the MH test investigates a violation of measurement invariance and summarises the result across all levels of the matching variable. Although measurement invariance results are not specific for each level of the matching variable, a matching variable is needed and its definition can influence the test results (Donoghue and Allen, 1993[48]).

Fifth, the MH test is based on frequentist hypothesis testing. Therefore, it is only possible to test whether the null hypothesis should be rejected. The quantification of evidence in favour of the null hypothesis (measurement invariance) is not possible. Furthermore, testing multiple hypotheses simultaneously is not possible. With MH, items can only be tested one by one, where it is assumed that the measurement invariant hypotheses across items can be tested independently from each other. In practice, the violation of measurement invariance can be related over items.

Sixth, the multigroup comparison is only possible under a common scale analysis. To test an item for a violation of measurement invariance at least one anchor or measurement invariant item is needed to identify the scale across groups. Otherwise, the MH test results cannot be interpreted in a meaningful way. It can be difficult to identify the anchors before starting the measurement invariance analysis.

Dorans and Holland (1993[49]) discussed a standardisation of the MH test (STAND), which is based on the principle that the expected performance of an item for each level of the matching variable is equal for the focal and reference group. The test does not require an equal number of responses across groups (i.e. unbalanced design), which is a requirement of the MH test. However, the other disadvantages of the MH test also apply to the standardised version.

Another class of methods for the detection of measurement non-invariance is based on a latent variable model (e.g. IRT model). A general procedure can be identified, where the fit of an IRT model in the focal group and reference group is compared. A measurement invariance test is focused on the significance of the difference between the item parameter estimates of both groups. An item is flagged as measurement variant, when the difference is significant. As an example, Thissen, Steinberg and Wainer (1993[50]) compared a compact model to an augmented model, which includes all parameters of the compact model as well as additional parameters that represent violations of measurement invariance. The significance of the additional model parameters is tested, where the compact model is hierarchically nested within the augmented model. The likelihood ratio test is used to evaluate whether the additional model parameters are significantly different from zero. The likelihood ratio test is assumed to be chi-squared distributed with the number of degrees of freedom equal to the number of item parameters differing between the reference and focal group. The null hypothesis is rejected for a large test statistic value. In a loglinear parameterisation of the model, the interaction parameter between group membership and item parameters is explicitly parameterised (Thissen, Steinberg and Wainer, 1993[50]).

One of the main disadvantages of this class of methods is that the effects, representing the violation of measurement invariance, are estimated, and measurement invariance tests are based on the estimates. For instance, the DIF effects of an item (i.e. the differences between item characteristic effects across groups) are usually difficult to estimate and often have large standard errors. The likelihood ratio test to evaluate the significance of the additional

parameters in the augmented model takes into account the low precision of the estimates. As a result, the (likelihood ratio) test is expected to have a low power, since it is based on estimates with low precision. Furthermore, low power can be expected for relatively small sample sizes. In this situation, the asymptotic sampling distribution of the test statistic might also not be appropriate leading to inaccurate test results. For a large sample size, small model deviations are easily significant. In that case, significant DIF effects might be detected, while in fact another misspecification of the model led to a significant value of the test statistic.

Multiple group confirmatory factor analysis (MGFCA) is a more advanced approach within this class of methods, since it also includes the specification of a population distribution. A chi-squared difference test can be too restrictive for large sample sizes and as an alternative, global model fit measures are used to evaluate the fit of the model under the measurement invariance assumption. Different fit indices are often used, since their sampling distributions are unknown and the exact behaviour of each index is not exactly known. The knowledge of the sampling distribution is critical to decide whether the estimated difference is indeed measurement non-invariance or sampling error. Another problem is that the sampling distribution is affected by the sample size, the characteristics of the model (e.g. number of items, factor structure) and the index that is used. Recommended cut-offs are not universally applicable and the same cut-off value can imply a different level of goodness of fit across different models (e.g. factor structures) and sample sizes.

Testing additional model parameters that represent a violation of measurement invariance cannot be used to examine the evidence in favour of measurement invariance. It can only be used to quantify the evidence in favour of an alternative. This makes the method not useful for examining the data evidence in favour of measurement invariance.

Finally, the additional model parameters in the augmented model need to be defined on a common scale and anchors (e.g. measurement invariant items) are needed to evaluate measurement invariance assumptions of the other items. Without knowing the anchors, a top-down or bottom-up procedure is needed, where (sets of) items are tested sequentially to identify possible violations of measurement invariance. However, the order of testing the items for partial or full measurement invariance influences the results and different procedures can lead to different results. There is a high probability of errors due to capitalisation on chance, when the significance levels are based on the assumption of independent hypotheses.

Score Purification Methods

In the MH procedure a matching variable (e.g. test score) is used to identify DIF across the levels of the matching variable. A common procedure in DIF detection is to purify the matching variable by excluding items that exhibit DIF. In an iterative purification method, the most significant item on a test according to the MH statistic is identified, and this item is omitted when computing the matching variable for the subsequent DIF analyses. The purification procedure can improve the power of the test but can also reduce the Type-1 error (Lee and Geisinger, 2015[51]).

Recently, different methods have been developed to purify latent variable scores in the presence of measurement variant items. These methods are aimed at producing comparable latent variable scores thereby accounting for the measurement variant items. In a Bayesian modelling framework approximate measurement invariance has been introduced, where the prior for the item parameters allows them to be different to some extent across groups. The

idea is that latent scores computed under the assumption of approximate measurement invariance are not contaminated by item bias. The items are allowed to function differently across groups. However, as mentioned earlier, the prior specification of the item parameters (e.g. intercepts and loadings) highly influences the results. The purification of the estimated scores depends on the specification of the priors, and usually the magnitude of the DIF for each item is unknown. The method is not very useful, when a priori no information is available about the size of the DIF.

The alignment method is another procedure that can be used to purify latent variable scores in the presence of DIF items. In the alignment method a rotation of the estimated solution is given, where the number of measurement invariant items is minimised. The estimated factor scores can be interpreted based on a subset of items that appear as measurement invariant.

The alignment method has some disadvantages, when the object is to identify the measurement invariant items. First, the method is not meant for testing the measurement invariance assumption of each item. Single items cannot be tested with this procedure, since the method is based on rescaling the solution of the estimated model. It can only provide a set of items that appear as measurement variant. Second, the identified items that show measurement non-invariance depend highly on the properties of the other items in the test. Items that exhibit a small to moderate level of measurement variance might not be detected, when other items exhibit a large level of measurement variance. Third, the method does not provide tools to formally test whether the identified discrepancies in the estimated item parameters are significant. It is not clear whether multiple hypothesis testing can be employed to identify (and/or the percentage of) the measurement variant and measurement invariant items.

Fourth, the rotated solution is obtained using a loss function, which assumes that there are always a few large variant items and many approximately invariant items. This approach is certainly not universal and might not be realistic for large-scale assessment data. Furthermore, different loss functions will lead to different results, but it is not clear which one will be the most suitable. In general, the optimisation criterion contains a priori information, which influences the final result, and this is beyond the control of the practitioner. Note that this loss function is also used in exploratory factor analysis (EFA), where it is desired to obtain subsets of small or large factor loadings. However, this approach does not translate nicely to measurement invariance testing, where optimisation criteria are more likely to differ across studies.

Fifth, the method breaks down when more items are measurement variant than measurement invariant. However, it is not possible to verify this, since usually it is not known which items exhibit measurement non-invariance. As stated by Asparouhov and Muthén (2014[19]), the alignment method returns the simplest model where most of the items are measurement invariant, this might however not be the optimal solution. They also showed that parameter biases increase with increasing levels of measurement non-invariance, decreasing group sizes and increasing number of groups.

Finally, the alignment method is restricted to multiple (fixed) groups (strata), and cannot be applied to the situation where the groups are also sampled from a larger population. The method can only be used to make inferences about differences across the groups in the sample.

Bayesian Hypothesis Testing of Measurement Invariance

The marginal measurement invariance approach proposed in this chapter is based on the Bayesian hypothesis testing framework, which has several advantages, and avoids several limitations of frequentist hypothesis testing. The main focus is on the posterior probability of a hypothesis, which represents the relative plausibility of the hypothesis given data and prior information. This posterior probability comprehends the data and prior information about the hypothesis. Without a preference for a specific hypothesis, the considered hypotheses are a priori equally likely.

The posterior probability of a null hypothesis H_0 given the data $\boldsymbol{y}$ can be expressed as $p(H_0|\boldsymbol{y})$. When an alternative hypothesis is considered, the posterior probability of H_0 is given by

$$p(H_0|\mathbf{y}) = \frac{p(\mathbf{y}|H_0)\,p(H_0)}{p(\mathbf{y}|H_0)\,p(H_0) + p(\mathbf{y}|H_1)\,p(H_1)}, \qquad \textbf{Equation 4.4}$$

where $p(H_0) = 1 - p(H_1)$ represents the prior probability of the null hypothesis. The ratio of posterior probabilities of the two hypotheses can be expressed as:

$$\frac{p(H_0|\mathbf{y})}{p(H_1|\mathbf{y})} = \frac{p(\mathbf{y}|H_0)}{p(\mathbf{y}|H_1)}\frac{p(H_0)}{p(H_1)} = BF_{01}\frac{p(H_0)}{p(H_1)} \qquad \textbf{Equation 4.5}$$

The left-hand side represents the posterior odds where the right-hand side contains the prior odds. The introduced Bayes factor (BF) represents the evidence from the data to update the prior beliefs. When more than two models are considered, each posterior probability of a model can be computed from the set of BFs for each pair of competing models. That is,

$$p(M_k|\mathbf{y}) = \frac{BF_{k0}}{\sum_{i=0}^{k} BF_{i0}} \qquad \textbf{Equation 4.6}$$

under the assumption of equal prior probabilities.

Although the Bayes factor (BF) (Kass and Raftery, 1995[52]) has important advantages, it has not been applied by many for evaluating assumptions of measurement invariance. The BF, following the hypothesis testing point of view of Jeffreys (1961[53]), allows one to evaluate evidence in favour of the null hypothesis. This is not possible in traditional significance testing, where the null hypothesis is not rejected, when evidence is lacking to decide otherwise. Failing to reject the hypothesis is of course no justification for using the model.

When using the BF, data can be used directly as evidence to give positive support to the null hypothesis. Besides data information, other external information can be used to evaluate the hypothesis. The BF makes it possible to directly assess the support in favour of the hypothesis. This is in contrast with the traditional way of testing, which is aimed at rejecting the null hypothesis. When different hypotheses are compared, the BF can be used

to identify the amount of support for each hypothesis given the data information (Kass and Raftery, 1995[52]).

The BF allows the comparison of nested and non-nested models, each representing different hypotheses. It allows the inclusion of different types of evidence (data and non-data information), since it is constructed from prior and posterior information. The interpretation of the BF is also straightforward. It represents a summary of evidence in favour of one of the models after observing all information, where each model represents a scientific theory or statistical hypothesis.

In BF testing the prior specifications are important, since the BF is computed by integrating over the model parameters. To make this more concrete, consider the BF in Equation 4.5, and let $\boldsymbol{\theta}$ denote the model parameters. Then, the BF can be represented as

$$BF_{01} = \frac{p(\mathbf{y}|H_0)}{p(\mathbf{y}|H_1)} = \frac{\int p(\mathbf{y}|\boldsymbol{\theta}, H_0)\, p(\boldsymbol{\theta}|H_0) d\boldsymbol{\theta}}{\int p(\mathbf{y}|\boldsymbol{\theta}, H_1)\, p(\boldsymbol{\theta}|H_1) d\boldsymbol{\theta}} \qquad \textbf{Equation 4.7}$$

which represents the ratio of marginal likelihoods (marginal distribution of the data). The marginal distribution of the data is obtained by integrating over the parameter space, where each model has its own prior specification for parameter $\boldsymbol{\theta}$. Hypotheses with order restrictions on the parameters of interest can be treated in the same way as hypotheses with equality constraints on the parameters of interest.

The BF scale is directly interpretable: substantial evidence of the null hypothesis is found when BF_{01} is between 3 and 10, and strong evidence between 10 and 30 (Jeffreys, 1961[53]; Kass and Raftery, 1995[52]). A Bayes factor outcome of greater than 3 indicates that there is substantial evidence in favour of the null hypothesis and that measurement invariance is plausible. When the Bayes factor is between 1/3 and 3, then there is no decisive evidence in favour of one of the hypothesis. The direct interpretation of the BF is appealing, and a sampling distribution of the BF is not needed to identify extreme BF outcomes.

There are more advantages of the BF. First, the BF automatically adjusts for model complexity and avoids overfitting. It will select the model that provides the best description of the data, while taking the complexity of the model into account. The complexity of the model is integrated in the prior distribution. A more complex prior will describe a larger part of the parameter space where the likelihood is relatively small and this will decrease the marginal likelihood (Hoijtink, 2012[54]).

Second, the BF can handle hypotheses that vary in complexity. For example, the BF can be used to compare a hypothesis of full measurement invariance to a hypothesis of partial measurement invariance, which will be represented by more model parameters.

Fractional Bayes Factor Testing

The integration over the prior probability in Equation 4.7 requires a proper prior distribution (i.e. a distribution that integrates to one). However, in evaluating measurement invariance hypotheses non-informative (improper) priors can be preferred to make data-based decisions. The construction of non-informative priors can be improved, when the priors do not need to be proper. A common strategy is to define a prior with the same functional form as the likelihood to obey any numerical restrictions on the parameter space, but without preferring any parameter value above another.

The Bayes factor does not provide interpretable outcomes when improper priors are used, since the outcomes will depend on the unknown normalising constants. However, a fractional Bayes factor (FBF) can be computed, which normalises the outcome by using a fraction of the data (O'Hagan, 1995[55]). The Bayes factor is computed as the probability of the data given the null hypothesis divided by the probability of the data given the alternative hypothesis:

$$\mathrm{BF}_{01} = \frac{p\left(\mathbf{y}|\mathrm{H}_0\right)}{p\left(\mathbf{y}|\mathrm{H}_1\right)}. \qquad \textbf{Equation 4.8}$$

The probabilities in the nominator and denominator are referred to as the marginal distribution of the data or marginal likelihood, given the model under the concerning hypothesis. The marginal likelihood given hypothesis $\mathrm{H_i}$ can be specified as follows (Raftery, 1995[56]):

$$p\left(\mathbf{y} \mid \mathrm{H_i}\right) = \int p\left(\mathbf{y} \mid \tau_k, \mathrm{H_i}\right) p\left(\tau_k \mid \mathrm{H_i}\right) d\tau_k \qquad \textbf{Equation 4.9}$$

where $\boldsymbol{\tau_k}$ stands for the parameters in the augmented model under hypothesis $\mathrm{H_i}$. To make objective decisions about the level of measurement variance, an improper prior is specified for the parameter $\boldsymbol{\tau_k}$, leading to an expression of the marginal distribution of the data up to an unknown constant.

A minimal information sample is used with the purpose of normalising the prior under the hypothesis. In order to take into account the improper prior, the marginal distribution of the data given the hypothesis is divided by a term, which serves as a normalising constant. The unspecified normalising constant is determined by integrating the prior times a fraction s of the likelihood of the data over the parameter space. The fraction s symbolises the minimal information needed to take into account the improper prior. The marginal distribution of the data under hypothesis H_i given fraction s can be represented as

$$p\left(\mathbf{y} \mid \mathrm{H_i}, s\right) = \frac{\int p\left(\mathbf{y} \mid \tau_k, \mathrm{H_i}\right) p\left(\tau_k \mid \mathrm{H_i}\right) d\tau_k}{\int p\left(\mathbf{y} \mid \tau_k, \mathrm{H_i}\right)^s p\left(\tau_k \mid \mathrm{H_i}\right) d\tau_k}, \qquad \textbf{Equation 4.10}$$

where $\boldsymbol{\tau_k}$ stands for the parameters in the model under hypothesis H_i. The denominator in Equation 4.10 represents a normalising constant to compensate for using an improper prior in the numerator. The interpretation of the resulting fractional Bayes factor remains the same as the interpretation of the Bayes factor. As a result, the advantages of BF testing also apply to the fractional Bayes factor testing.

The (fractional) Bayes factor test supports investigating all items simultaneously on violations of measurement invariance. Simultaneously testing multiple hypotheses is rather straightforward, since it is also based on comparing the marginal distributions of the data. Traditional measurement invariance tests in IRT and factor analysis differ in this respect, where the significance level of each single hypothesis test need to be restricted to compensate for the number of comparisons that are made (i.e. a multiple testing correction).

Posterior Predictive Testing

Up until now, a universal Bayesian method to evaluate competing models does not exist. Various methods have been proposed. One of the most popular Bayesian method is posterior predictive model assessment, where a discrepancy measure is defined to evaluate a model assumption (Gelman et al., 2003[57]). The method is straightforward, intuitive, and can be applied to assess the fit of different aspects of the model. Posterior predictive checks have been proposed to evaluate approximate measurement invariance.

Although posterior predictive checks are easily implemented in various ways, the method has some drawbacks. First, the statistical significance of an event is based on a posterior predictive p-value, which might not be uniformly distributed under the posited model. This relates to fact that in the procedure the data are used twice, which leads to conservative behaviour of the posterior predictive p-value and can fail to detect model misfit. As a result, when using posterior predictive p-values to test a null hypothesis, the empirical Type I error rates are often below nominal values (Levy, Mislevy and Sinharay, 2009[58]). Therefore, posterior predictive model assessment is often viewed as a diagnostic measure to identify possible misfits, instead of a formal test for model misfit (Gelman et al., 2003[57]). This relates to the idea to evaluate model fit by collecting pieces of statistical evidence in combination with substantive theory.

Second, a discrepancy measure is often not completely targeted for the specific model misfit. For example, the odds ratio (Levy, Mislevy and Sinharay, 2009[58]) has been proposed as a discrepancy measure to detect violations of local independence. However, the odds ratio is a measure of association for paired observations and does not represent directly the assumption of local independence. Without an accurate representation of the model misfit by the discrepancy measure, significant results might be caused by violating another model assumption, and incorrect inferences might be drawn.

Third, the posterior predictive check is mainly used to evaluate the compatibility of the model with the data and no fully specified alternative model is available. However, when competing theories are investigated, and alternative hypotheses are present, the posterior predictive check is not suitable to evaluate the competing hypotheses. A problem arises when different models, apparently reasonable in comparison to the data, lead to different results.

Marginal Random Item Effects Model

The marginal measurement invariance test approach is based on detecting a correlation among responses to the same item within each group, without conditioning on group-specific item effects. In the MH test, item bias (the interaction between the group and the item) is examined by investigating whether the odds of getting an item correct significantly differ between groups, given the level of the matching variable. For a measurement invariant item, the sample data are independently distributed (i.e. a simple random sample) for each level of the matching variable. For a measurement variant item, the sample data are a cluster sample or a stratified sample, since the odds differ across groups. The marginal measurement invariant test is also aimed at testing the correlation among responses within each group, where a significant positive correlation implies a violation of measurement invariance. However, the item bias is not estimated only by the within-group correlation.

In the remainder of this chapter, the marginal random item effects model is introduced, and it is explained how it can be used to test measurement invariance. The fractional Bayes factor is used to objectively compare competing hypotheses to accommodate an improper

prior for the implied degree of measurement variance. Results of a simulation study are given, where the functioning of the fractional Bayes factor was compared to that of the posterior predictive check based on the Mantel-Haenszel chi-square statistic χ^2_{MH}, since the latter is a commonly used tool to detect measurement non-invariance (Holland and Wainer, 1993[4]). Extensions of the method are discussed in order to make it applicable to randomly selected groups, for which parameter recovery is also evaluated in a simulation study. Finally, the method is applied to empirical data using data from the European Social Survey (ESS).

To explain the marginal test approach, the conditional modelling approach where group-specific item parameters are modelled serves as a reference. For instance, the random item effects model is a conditional model in which a normal distribution is assumed for the group-specific item parameters. This model has been used by Verhagen and Fox (2013[40]) and De Jong, Steenkamp and Fox (2007[45]) to detect violations of measurement invariance. In this chapter, it is shown that a marginal model can be derived from this random item effects model such that group-specific item parameters no longer need to be modelled. The marginal test approach avoids estimating the differential effect of item functioning and, subsequently, testing the significance of the effects, since this approach will lead to a loss in power. The one-parameter multilevel IRT model will be used for illustration, as described by Bock and Zimowski (1997[59]) and Azevedo, Andrade and Fox (2012[60]), and the probability of answering an item correctly is given by

$$P(Y_{ijk}=1\,|\,\theta_{ij},b_k)=\Phi(\theta_{ij}-b_k), \qquad \textbf{Equation 4.11}$$

where θ_{ij} is the underlying ability of person i in group j, and b_k is the difficulty of item k and $\Phi(\cdot)$ the normal cumulative distribution function. Parameter b_k reflects the required value of the underlying ability θ in order for the test taker to have an expected probability of .5 of answering the item correctly.

The Random Item Effects Model

To illustrate the method, the one-parameter multilevel IRT model is used and in the discussion the generalisation of the method to other IRT models is discussed. Before turning to the one-parameter multilevel IRT model, assume continuous responses to items, symbolised by Z_{ijk} In a random item effects model, this latent response variable is modeled as follows:

$$Z_{ijk}=\theta_{ij}-b_{jk}+\varepsilon_{ijk},\varepsilon_{ijk} \sim N(0,1), \qquad \textbf{Equation 4.12}$$

where

$$b_{jk}=b_k+\varepsilon_{jk},\varepsilon_{jk}\sim N(0,\tau_k) \qquad \textbf{Equation 4.13}$$

and

$$\theta_{ij}=\mu_{\theta_j}+r_{ij},r_{ij}\sim N\left(0,\sigma^2_\theta\right). \qquad \textbf{Equation 4.14}$$

In Equation 4.12, the random item effects model is shown, where the latent response variable Z_{ijk} is independently and identically distributed given the group-specific item difficulty parameter b_{jk} and person parameter θ_{ij}. As illustrated in Equation 4.13, the random item effects parameter is assumed to be normally distributed with the mean equal to the invariant item difficulty parameter b_k and variance τ_k. The variance parameter τ_k stands for the between-group variance with respect to the random item difficulty parameter, and it represents the degree of measurement variance. The person parameter is assumed to be normally distributed with a group-specific mean.

The Marginal Modelling Approach

The random item effects model can be marginalised by integrating out the group-specific item parameters. This can be done by plugging Equation 4.13 into Equation 4.12. It follows that

$$\begin{aligned} Z_{ijk} &= \theta_{ij} - b_{jk} + \varepsilon_{ijk} \\ &= \theta_{ij} - b_k + \varepsilon_{ijk} + \varepsilon_{jk} \\ &= \theta_{ij} - b_k + E_{ijk}, \end{aligned} \qquad \textbf{Equation 4.15}$$

where the errors for item k ($\boldsymbol{E}_k$) are assumed to have a multivariate normal distribution with a mean of zero and covariance matrix $\boldsymbol{\Sigma}_{\mathrm{k}}$

In this marginal model, the latent response variable no longer depends on the group-specific item parameter b_{jk} and one item difficulty parameter b_k (i.e. the measurement invariant item difficulty) applies to all groups. As a result, the degree of measurement variance is included in the error term. Note that in this marginal model, conditional independence no longer applies due to the fact that group-specific item parameters are not specified. In the marginal random item effects model as described here, Z_{ijk} has a multivariate normal distribution. The presence of measurement non-invariance is absorbed into the covariance structure of the error term.

To explain the covariance structure of the marginal model, in Equation 4.15, in more detail, let the measurement error be normally distributed $\varepsilon_{ijk} \sim N(0, \sigma_{sk}^2)$. Then, covariance matrix $\boldsymbol{\Sigma_k}$ can be specified. In the first case, $i = i'$, which automatically implies that $j = j'$. This reflects the covariance of two responses of person i in group j In the second case, $i \neq i'$ but $j = j'$. That is, different persons i and i' are in the same group j. The third case consists of the covariance of different persons i and i' in different groups j and j'. It can be concluded that the (co)variances in these three different cases are equal to $\tau_k + \sigma_{\varepsilon k}^2$, τ_k, and 0, respectively:

$$\begin{aligned} \boldsymbol{\Sigma}_k &= \operatorname{cov}\left(\varepsilon_{jk} + \varepsilon_{ijk}, \varepsilon_{j'k'} + \varepsilon_{i'j'k}\right) \\ &= \operatorname{cov}\left(\varepsilon_{jk}, \varepsilon_{j'k}\right) + \operatorname{cov}\left(\varepsilon_{ijk}, \varepsilon_{i'j'k}\right) \\ &= \begin{cases} \operatorname{var}(\varepsilon_{jk}) + \operatorname{var}(\varepsilon_{ijk}) = \tau_k + \sigma_{\varepsilon k}^2 & \text{if } i = i', j = j' \\ \operatorname{var}(\varepsilon_{jk}) = \tau_k & \text{if } i \neq i', j = j' \\ 0 & \text{if } j \neq j'. \end{cases} \end{aligned} \qquad \textbf{Equation 4.16}$$

In this marginal model, there is only one item difficulty parameter present, which applies to all the groups, instead of there being an item difficulty parameter for each group separately. The possible error due to measurement non-invariance is no longer explicitly modelled in b_k but included in the covariance structure of the error distribution. In $\boldsymbol{\Sigma_k}$ the presence of measurement non-invariance is captured by the covariance of different observations within a group, specified by the second case in Equation 4.16. It is proposed that in order to test whether measurement non-invariance is present (and to what degree), one should evaluate τ_k.

When the groups are randomly selected from a population, the covariance structure for the responses to item *k* in group *j* is given by

$$\boldsymbol{\Sigma}_{jk} = \sigma^2_{\varepsilon k}\mathbf{I}_m + \tau_k\mathbf{J}_m, \qquad \textbf{Equation 4.17}$$

where $\mathbf{I}_m$ is the identity matrix and $\mathbf{J}_m$ a matrix of ones; *m* stands for the number of observations in each group, and equal group sizes (balanced design) are assumed. In the covariance structure of Equation 4.17, parameters τ_k on the off-diagonal positions represent the implied covariance between latent responses due to the clustering of responses in groups. Parameters τ_k on the diagonal positions contribute to the variance in item difficulty across groups. For randomly selected groups, the random item effect parameter is used to model the clustering of responses in groups as well as the variability in item functioning across groups. The groups are sampled from a population, and the random item effects variance represents the variance in item functioning in the population of groups. For all items *k*, when binary response data are observed, the variance parameter $\sigma^2_{\varepsilon k}$ will be fixed to one to identify the scale.

For a fixed number of groups, there is no population distribution defined for the groups. The considered groups are of specific interest, and the object is to evaluate the assumption of measurement invariance for the selected groups. In the parametrisation presented in Equation 4.17, the covariance parameters can modify the covariance between response observations as well as the total amount of variance in response observations. Therefore, a different parametrisation is considered to avoid the situation that the covariance parameter, τ_k, can also represent variability in item functioning across groups. For a fixed number of groups, the diagonal components in the covariance matrix $\boldsymbol{\Sigma}_{jk}$ should only represent measurement error variance and not also variation in item functioning in the population. Therefore, to restrict the additional contribution of the covariance components to the total variance, the variance parameter equals $\sigma^2_{\varepsilon k} = 1 - \tau_k$. In that case the total variance, represented by the diagonal terms, is always equal to one, and the covariance components are not allowed to increase the total variance. This leads to the following covariance matrix of the error terms for each group *j* and item *k*,

$$\boldsymbol{\Sigma}_{jk} = \left(1-\tau_k\right)\mathbf{I}_m + \tau_k\mathbf{J}_m \qquad \textbf{Equation 4.18}$$

It follows that the values on the diagonal are equal to 1 and the off-diagonal values are equal to τ_k. In this covariance structure $\boldsymbol{\Sigma}_{jk}$ the τ_k is a correlation parameter, since the diagonal consists of ones.

The marginal random item effects model for binary response data is represented by a generalised multivariate probit model:

$$P\left(\mathbf{Y}_{jk}=\mathbf{y}_{jk}\left|\boldsymbol{\theta}_{j},\mathbf{b}_{k},\boldsymbol{\Sigma}_{jk}\right.\right)=\int_{\Omega(\mathbf{y}_{jk})}\Phi\left(\mathbf{z}_{jk}\left|\boldsymbol{\theta}_{j},\mathbf{b}_{k},\boldsymbol{\Sigma}_{jk}\right.\right)d\mathbf{z}_{jk} \qquad \textbf{Equation 4.19}$$

where the latent response data is truncated multivariate normally distributed according to the set $\Omega_{jk}=\{\boldsymbol{z}_{jk}\colon z_{ijk}\leq 0 \text{ if } y_{ijk}=0, z_{ijk}\geq 0 \text{ if } y_{ijk}=1\}$. The person parameters are assumed to be normally distributed according to Equation 4.14, where a non-informative Jeffreys prior is specified for the hyperparameters $(\mu_{\theta_i},\sigma_\theta^2)$, $p(\mu_{\theta_i},\sigma_\theta^2)\propto\sigma_\theta^{-2}$. A normal distribution is assumed for the invariant item parameters,

$$b_k\sim N(\mu_b,\sigma_b^2) \qquad \textbf{Equation 4.20}$$

and a normal-inverse gamma prior is specified for the hyperparameters,

$$\begin{aligned}\mu_b &\sim N(\mu_0,\sigma_b^2/n_0),\\ \sigma_b^2 &\sim IG(\alpha_b,\beta_b),\end{aligned} \qquad \textbf{Equation 4.21}$$

where the $n_0\geq 0$ determines the weight of the prior specification of μ_b.

Prior and Posterior of the Covariance Parameter

To make inferences about the dependency structure under the marginal random item effects model, interest is focused on the covariance parameter. A non-informative prior can be defined, which takes on the same functional form as the likelihood. This non-informative prior for τ_k is given by

$$p(\tau_k)\propto(1/m+\tau_k)^{-1}, \qquad \textbf{Equation 4.22}$$

where *m* represents the number of response observations in each group. The covariance parameter τ_k is restricted to be greater than *-1/m*, which follows from the fact that the covariance matrix needs to be invertible (i.e. positive definite). The non-informative prior does not prefer any parameter value above any other. Furthermore, the parameter value $\tau_k=0$ is not on the boundary of the parameter space. The value of zero for the covariance parameter means that the item is measurement invariant, since it does not induce an additional correlation between the responses within each group. With $\tau_k=0$, there is no support for a random item effect, and the responses to item *k* are independently distributed given the measurement invariant item difficulty and the person parameter.

A positive covariance parameter represents a common covariance between the responses to item *k* within each group, while conditioning on the person parameter and the item difficulty parameter b_k. This clustering effect can be represented by a random item effects parameter, and it represents a violation of measurement invariance.

When the covariance parameter is negative, there is even less variation in the average item scores of item *k* across groups than the variation in average item scores for a measurement invariant item *k* where $\tau_k=0$. This means that the sample heterogeneity among groups would be even lower than the one for $\tau_k=0$, when there is no heterogeneity across groups. The marginal random item effects model, after integrating out the random item effect,

represents a wider parameter space for the covariance parameter, including a negative support. This property proves to be beneficial for constructing conditionally conjugate priors for τ_k.

Another beneficial effect of the prior is that the posterior distribution for the covariance parameter can be obtained in closed form. This property is used to define a (fractional) BF test for evaluating the dependency structure. In Annex 4.A, the analytical distribution of the covariance parameter is given.

In order to estimate the parameters of the marginal random item effects model, a Markov chain Monte Carlo (MCMC) algorithm can be used. The MCMC algorithm can be found in Fox et al. (2017[42]).

Fractional Bayes Factor

In Annex 4.B, analytical expressions of the fractional Bayes factors are given to evaluate measurement invariance hypotheses. For randomly selected groups and for fixed groups, fractional Bayes factors are computed to evaluate $H_0: \tau_k = 0$ and $H_u: \tau_k \neq 0$, which represents the null hypothesis that the item is measurement invariant and the alternative (unrestricted) hypothesis that the item is not measurement invariant, respectively. For an item that cannot be characterised as measurement invariant, it is possible that (a) the item is measurement variant and shows differential item function across groups (i.e. item responses are group-specific positively correlated), or (b) the item does not contribute to the measurement scale (item responses are group-specific negatively correlated).

The fractional Bayes factor, denoted as FBF_{0u} evaluates the evidence in favour of measurement invariance ($H_0: \tau_k = 0$) against the hypothesis that there is no measurement invariance $H_u: \tau_k \neq 0$. Another fractional Bayes factor is considered and referred to as FBF_{02}, which evaluates the evidence in favour of measurement invariance $H_0: \tau_k = 0$ against the alternative hypothesis that there is measurement non-invariance $H_2: \tau_k > 0$, such that a negative τ_k is not supported by either the null hypothesis or the alternative hypothesis. In this case, the data is used to evaluate the evidence in favour of measurement invariance or in favour of measurement non-invariance.

Simulation Study for Stratified Groups

In this simulation study, parameter recovery of the marginal random item effects model was evaluated for the situation of a fixed number of groups (strata). The first goal of this simulation study was to test whether the marginal model is able to accurately estimate the degree of measurement variance. The second goal of this simulation study was to evaluate the use of the fractional Bayes factor to decide whether or not the degree of measurement variance in an item is equal to zero.

The fractional Bayes factor was used to accommodate for the improper prior for the measurement non-invariance parameter τ_k. This improper prior assumes a uniform distribution for the possible degrees of measurement non-invariance, which makes it possible to objectively evaluate the measurement invariance assumption. The fractional Bayes factor approach has several important advantages. First, it is able to test for measurement non-invariance in all of the items simultaneously and does not require a sequential test procedure in which items are tested one by one. Second, anchor items are not needed, and full measurement invariance can also be tested using the same procedure. Third, it takes into account both the null hypothesis (H_0), which states that measurement

invariance holds, as well as the alternative hypothesis (H_u), which states that measurement invariance does not hold.

The posterior predictive *p*-value based on the Mantel-Haenszel χ^2_{MH} statistic (ppp χ^2_{MH}) does not have these advantages, but the functioning of the ppp χ^2_{MH} is compared to the functioning of the fractional Bayes factor. The χ^2_{MH} statistic was used as a discrepancy measure in a posterior predictive check in order to evaluate the measurement invariance assumption. Assume the persons were divided over subgroups based on their total test score. In a test with 10 items, this entails that the total number of subgroups G = 11, since a total score from 0 to 10 is possible. Since the χ^2_{MH} statistic needs anchor items, all the other items of the test are chosen as anchor items. The object was to identify which items were measurement invariant. Here, a conservative approach was followed, where each item was tested by assuming the other items to be measurement invariant. In practice, it is usually not known which items are measurement invariant, and so an assumption needs to be made in order to test a single item.

Further details of the simulation study can be found in Annex 4.C. Table 4.1 presents the results of the simulation study. In this simulation, measurement non-invariance increases across items. Parameter τ represents the simulated degree of measurement variance whereas τ' represents the estimated degree of measurement variance by the posterior mean. Column $\tau - \tau'$ shows the difference between the simulated measurement non-invariance and the estimated measurement non-invariance by the posterior mean computed under the marginal model. Table 4.1 shows that the estimated degree of measurement variance τ' differs a maximum of .089 from the simulated degree of measurement variance τ. The smallest absolute difference between the two values is equal to .001. It appears here that when the degree of measurement variance is smaller than .075, the posterior mean, as a point estimate, tends to overestimate the degree of measurement variance. When the degree of measurement variance is greater than .100, it tends to underestimate the degree of measurement variance.

Table 4.1. Fixed groups: Results of the simulation study for estimating the degree of measurement variance

Based on 50 replications

				Fractional Bayes Factor				Posterior Predictive Check		
Item	τ	τ'	$\tau - \tau'$	ln(FBF_{0u})	FBF_{0u}	ln(FBF_{02})	FBF_{02}	ppp χ^2_{MH}	Range ppp χ^2_{MH}	%ppp < 0.05
1	−0.002	0.033	−0.035	−0.755	0.470	−0.158	0.853	0.282	[0.000, 0.908]	28
2	0.000	0.038	−0.038	−1.208	0.299	−0.646	0.524	0.309	[0.000, 0.911]	36
3	0.025	0.047	−0.022	−2.463	0.085	−2.009	0.134	0.251	[0.000, 0.984]	44
4	0.050	0.077	−0.027	−7.980	<0.001	−7.686	<0.001	0.129	[0.000, 0.919]	60
5	0.075	0.076	−0.001	−7.036	0.001	−6.803	0.001	0.094	[0.000, 0.717]	68
6	0.100	0.097	0.003	−13.659	<0.001	−13.415	<0.001	0.111	[0.000, 0.776]	70
7	0.125	0.095	0.030	−11.875	<0.001	−11.703	<0.001	0.070	[0.000, 0.908]	80
8	0.150	0.099	0.051	−13.128	<0.001	−12.927	<0.001	0.100	[0.000, 0.869]	76
9	0.175	0.114	0.061	−18.977	<0.001	−18.791	<0.001	0.075	[0.000, 0.833]	78
10	0.200	0.111	0.089	−18.864	<0.001	−18.711	<0.001	0.068	[0.000, 0.890]	80

Notes: FBF_{01}: fractional Bayes factor, where H_0:τ =0 and H_u: $\tau \neq 0$; FBF_{02} = fractional Bayes factor, where H_0: τ=0 and H_2: τ>0; ppp: posterior predictive p-value based on the Mantel-Haenszel χ^2_{MH} statistic; ppp χ^2_{MH}: mean of the posterior predictive p-values over the 50 replications; Range ppp χ^2_{MH}: range of the found posterior predictive p-values over the 50 replications; %ppp < 0.05 shows the percentage of the 50 replications that resulted in a posterior predictive p-value based on ppp χ^2_{MH}< .05.

The posterior mean estimate of the variance parameter differs from the mode, since the posterior distribution is skewed. For a small (large) variance parameter, the posterior is skewed to the right (left) and the mean is higher (lower) than the posterior mode and over- (under) estimates the true value. For this reason, evaluating the presence of measurement non-invariance using point estimates is not recommended. In order to test whether the estimated degree of measurement variance $\tau' = 0$ or not, a fractional Bayes factor is computed and compared to the functioning of the ppp χ^2_{MH} for model selection. To compute the fractional Bayes factor, posterior samples are used, not point estimates, because posterior samples take the skewness of the posterior into account.

Under the marginal random item effects model, the natural logarithm of the fractional Bayes factor is computed (see Table 4.1). These results can be found under columns labelled $\ln(FBF_{0u})$ and $\ln(FBF_{02})$, which show that the higher the degree of measurement variance, the more negative the natural logarithm of the fractional Bayes factor. Table 4.1 also shows the fractional Bayes factors (columns FBF_{0u} and FBF_{02}). Though FBF_{02} performs very well for all the items, it is greater than 1/3 for items 1 and 2. Therefore, it can be concluded that items 1 and 2 are measurement invariant. However, FBF_{0u} shows support for the alternative hypothesis (H_u) for item 2, while this actually is a measurement invariant item. So, FBF_{02} performs better than FBF_{0u} in deciding whether item 2 is measurement invariant. This can be explained as follows. FBF_{02} results for items 1 and 2 show more support for the measurement invariance hypothesis (H_0) than FBF_{0u}, since the alternative hypothesis is restricted to the measurement non-variance hypothesis (H_2). Therefore, support for small, negative values of τ_k do not contribute to evidence in favour of alternative hypothesis H_2 ($\tau_k > 0$), whereas those values do contribute to alternative hypothesis H_u ($\tau_k \neq 0$). So, more power was obtained in detecting measurement invariance by restricting the alternative hypothesis to measurement non-variance (H_2).

Hypotheses H_0 and H_u were equally likely for items 1 and 2, but the fractional Bayes factors were not equal to one. The alternative hypothesis H_u also covers τ values, which are close to, but not exactly equal to, zero. The data give the most support to τ values equal to or close to zero, which makes H_u slightly more attractive than H_0.

For items 3-10, the fractional Bayes factors indicate that measurement non-invariance is present, and this was also simulated for these items. However, note that alternative hypothesis H_2 represents measurement non-invariance, whereas the evidence in favour of alternative hypothesis H_u represents all values of $\tau_k \neq 0$. For instance, for item 3, it is 11.76 times (1/0.85) more likely that $\tau_3 \neq 0$, but only 7.46 times (1/.134) more likely that $\tau_3 > 0$, which represents measurement non-invariance.

The results of the ppp χ^2_{MH} can be found in the last three columns of Table 4.1. It is hard to draw conclusions based on these values, since there is no common (universal) cut-off score. A ppp χ^2_{MH} close to zero shows discrepancies between the model that assumes measurement invariance and the observed data. Items 1-3 appear to have a smaller degree of discrepancy; items 4-10 appear to have a larger degree of discrepancy, since these values are closer to zero. This result is not exactly in line with that for the simulated data, since measurement non-variance was also present in item 3, which could not be clearly concluded from the results of the ppp χ^2_{MH}. Column %ppp < 0.05 shows the percentage of the 50 replications in which ppp χ^2_{MH} values were extreme (i.e. close to zero). Here, ppp χ^2_{MH} values are interpreted as extreme when they are less than .05. This column is provided to offer more insight with respect to the distribution of the ppp χ^2_{MH}. It is not meant as a threshold value for either accepting or rejecting the model. From this column, it can be

concluded that for items 1-3, ppp $\chi^2_{MH} < .05$ for less than half of the 50 replications, and for items 4-10, ppp $\chi^2_{MH} < .05$ for more than half of the replications.

The fractional Bayes factor has considerable benefits compared to the ppp χ^2_{MH} statistic. First, the fractional Bayes factor is able to test the degree of measurement variance for all the items at once, without the need to specify anchor items. Second, it compares the probability of the data given the null hypothesis to the probability of the data given the alternative hypothesis. This entails that both hypotheses are evaluated and the degree of support for each of them is compared. Consequently, the results are easy to interpret, since they either provide a preference for one of the two models or indicate that there is no preferable model. Finally, unlike the χ^2_{MH} statistic, which is only applicable for the comparison of two groups, the fractional Bayes factor can be computed for two or more groups. As expected, the results of the fractional Bayes factor are more convincing compared to those of the χ^2_{MH} statistic. Together with the other benefits of the fractional Bayes factor, it appears that this is an improved tool for detecting the presence of measurement non-invariance.

Simulation Study for Sampled Groups

In a second simulation study, parameter recovery by the marginal random item effects model was evaluated in the situation where groups are randomly selected from a larger population. In that case, test results about measurement invariance can be generalised to the population of groups from which the sample was taken. The covariance structure defined in Equation 4.17 was assumed for the responses to item k. The corresponding marginal random item effects model was tested by estimating the degree of measurement variance τ_k for every item; $\sigma^2_{\varepsilon k} = 1$ to identify the scale. Furthermore, the fractional Bayes factor was used to quantify the evidence against the hypothesis that the degree of measurement variance was equal to zero.

Table 4.2 presents the results of the simulation study. In this simulation, measurement non-invariance increases across items as in the previous simulation. The same symbols are used, where τ represents the simulated degree of measurement variance and τ' represents the estimated measurement variance. Column $\tau - \tau'$ shows the difference between the simulated degree of measurement variance and the estimated degree of measurement variance. The estimated degree of measurement variance τ' differs only a small amount from the simulated degree of measurement variance τ. The smallest difference is 0.000; the greatest absolute difference is 0.032. There appears to be an overestimation of the degree of measurement variance when the difference is less than 0.075 and an underestimation of the degree of measurement variance when the difference is greater than 0.125. However, this underestimation and overestimation is present to a lesser extent compared to the estimates for the fixed number of groups (see section titled Simulation Study for Fixed Groups). In this case, the variance in item responses between groups is also used to estimate τ. Again, the posterior mean will overestimate the true value when the posterior distribution is right-skewed and underestimate the true value when it is left-skewed.

In order to test whether $\tau = 0$ (H_0) or $\tau \neq 0$ (H_u), fractional Bayes factor FBF_{0u} was computed. When looking at the natural logarithm of the fractional Bayes factor in column $\ln(FBF_{0u})$ in Table 4.2, it can be seen that the greater the simulated degree of measurement variance τ gets, the more negative the natural logarithm of the fractional Bayes factor becomes. The FBF_{0u} results show correctly that for items 1 and 2, there is more support for the null hypothesis (H_0), and for items 3 and 10 there is substantially more support for the

alternative hypothesis (H_u). When looking at the results of FBF_{02} in the last column of Table 4.2, where the alternative hypothesis is restricted to measurement non-invariance (H_2: $\tau > 0$), it can be seen that there is a large degree of support for the measurement invariance hypothesis for items 1 and 2. For item 3, it is approximately 5.13 times more likely that the item is measurement variant than that it is measurement invariant. The results show that items 4-10 are measurement variant. It can be concluded that the results are in line with the simulation parameters used to generate the data.

Compared to the simulation study with stratified groups, the fractional Bayes factor results for items 1 and 2 show more support for the measurement invariance hypothesis. There is more information in the data about the exact value of the parameter, since both within- and between-group variation is used. For fixed (stratified) groups, only within-group information is used. As a result, estimates for the degree of measurement variance for randomly selected groups is more accurate compared to estimates for the degree of measurement variance for fixed groups.

Furthermore, for items 1 and 2, there is more data evidence in favour of the null hypothesis (H_0), which makes the alternative hypothesis (H_u) less attractive. Note that the data still provides some support for values near zero, which makes H_u slightly more attractive than H_0, leading to a fractional Bayes factor $FBF_{0u} < 1$ for item 2. However, when the alternative hypothesis (H_2) is considered, then small negative values of τ do not contribute to the evidence against the null hypothesis.

Table 4.2. Sampled groups: Results of the simulation study for estimating the degree of measurement variance

Based on 50 replications

Item	τ	τ'	$\tau - \tau'$	ln(FBF_{0u})	FBF_{0u}	ln(FBF_{02})	FBF_{02}
1	−0.020	0.002	−0.022	0.449	1.566	4.641	103.607
2	0.000	0.013	−0.013	−0.471	0.625	2.971	19.506
3	0.025	0.036	−0.011	−4.520	0.011	−1.636	0.195
4	0.050	0.055	−0.005	−9.415	<0.001	−6.594	0.001
5	0.075	0.075	0.000	−15.436	<0.001	−12.554	<0.001
6	0.100	0.100	0.000	−22.808	<0.001	−19.975	<0.001
7	0.125	0.125	0.000	−31.127	<0.001	−28.283	<0.001
8	0.150	0.140	0.010	−36.227	<0.001	−33.376	<0.001
9	0.175	0.143	0.032	−37.435	<0.001	−34.570	<0.001
10	0.200	0.188	0.012	−53.049	<0.001	−50.185	<0.001

Notes: FBF_{01}: fractional Bayes factor, where H_0:τ =0 and H_u: $\tau \neq 0$; FBF_{02} = fractional Bayes factor, where H_0: τ=0 and H_2: τ>0.

Evaluating Measurement Invariance Assumptions of the European Social Survey Items

There are many areas where methods for the detection of measurement non-invariance can be useful. International surveys, in which the answers of respondents across countries are compared, are one such example. To demonstrate the application of the fractional Bayes factor under a marginal random item effect model for detecting measurement non-invariance, data from the European Social Survey (ESS) round 7 (year 2014) was used.

Currently, the developed software for marginal measurement invariance testing is limited to equal group sizes (balanced design) and binary response data. However, the marginal

test procedure can be generalised to the more general situation of unbalanced groups and mixed (e.g. continuous, dichotomous, polytomous) response data, which is explained in the discussion. To illustrate the method, a balanced random sample was drawn from the countries included in this empirical example, where 1 750 response observations were sampled from each country. In the unbalanced data set the number of observations ranged from 1 769 to 2 390. The response data were also dichotomised. The possible answers to each of the included questions are on a scale from 0 to 10, as illustrated in Table 4.3. In items 1-3 and item 6-8, 0 stands for a negative attitude towards immigrants and 10 stands for a positive attitude towards immigrants. The five most negative categories towards immigrants (categories 0-4) were coded as 1 and the other six categories (5-10) which reflect (relatively) positive attitudes about immigrants were coded as 0. For items 4 and 5, 0 stands for a positive attitude towards immigrants and 10 stands for a negative attitude towards immigrants. The five most negative categories with respect to attitude towards immigrants (categories 6-10) were coded as 1 and the other six categories (categories 0-5) which reflect (relatively) positive attitudes towards immigrants were coded as 0. Note that the dichotomisation can influence the test results, when for instance an item exhibits measurement non-invariance for one of the response categories but not homogenously across response categories.

When measurement non-invariance is present, people who have the same attitude towards immigrants have a different probability of giving the same answer depending on their country. Otherwise stated, for a measurement variant item, respondents who have the same attitudes but are from different countries have unequal probabilities of scoring positively towards immigrants. In order to show the application of the model to empirical data, eight items were selected from the ESS survey. The eight items selected (Table 4.3) concerned the topic of immigration, since it was likely that measurement non-invariance is present in items such as these. The items contributed to the same scale, which measures attitude towards immigrants. Measurement invariance was tested for two different situations: a fixed number of groups (stratified groups) and randomly selected groups. In the situation of fixed groups, the goal was to make inferences about the degree of measurement variance between two selected countries, in this case Belgium and Sweden. The number of observations included was 1 750 for each country. So, the total number of observations was 3 500.

Table 4.3. European Social Survey items selected for the application study

Item	Statement	Response Scale
1	Immigrants generally take jobs away or help to create new jobs	0 Take jobs away – 10 Create new jobs
2	Immigrants take out more than they put in regarding taxes and welfare or not	0 Generally take out more – 10 Generally put in more
3	Immigrants make country's crime problems worse or better	0 Crime problems made worse – 10 Crime problems made better
4	Mind if immigrant of different race or ethnic group was your boss	0 Not mind at all – 10 Mind a lot
5	Mind if immigrant of different race or ethnic group would marry close relative	0 Not mind at all – 10 Mind a lot
6	The country's cultural life is undermined or enriched by immigrants	0 Cultural life undermined – 10 Cultural life enriched
7	Immigration is bad or good for country's economy	0 Bad for the economy – 10 Good for the economy
8	Immigrants make the country a worse or better place to live	0 Worse place to live – 10 Better place to live

Source: European Social Survey (2015), ESS-7 2014 documentation report. Edition 1.0. Bergen, European Social Survey Data Archive, Norwegian Social Science Data Services for ESS ERIC.

In the situation of randomly selected groups, six countries were selected and included in the study, and the goal was to investigate measurement invariance assumptions for items across the countries included in the ESS. It was assumed that the six countries (i.e. Austria, Belgium, Czech Republic, Denmark, Germany and Switzerland) well represented the ESS countries. The number of observations included was 1 500 for each country, for a total number of observations of 9 000. In the current model, additional sampling weights were not taken into account. Therefore, it is possible that the empirical results were affected by exclusion of the weights. In the discussion section, different options for dealing with survey weights are discussed.

In order to decide whether or not measurement non-invariance was present in an item, fractional Bayes factors were computed. As in the simulation study, the FBF_{0u} represents the fractional Bayes factor to evaluate the evidence in favour of measurement invariance (H_0: $\tau = 0$) compared to no measurement invariance (H_u: $\tau \neq 0$). The FBF_{02} represents the evidence in favour of measurement invariance compared to measurement non-invariance (H_2: $\tau > 0$).

Table 4.4 shows the results for the situation where the degree of measurement variance is estimated for both fixed and randomly selected groups. First, the results for the degree of measurement variance for a fixed number of groups (i.e. Belgium and Sweden) are discussed. Results for the fractional Bayes factors FBF_{0u} and FBF_{02} are presented in this table as well as the results for the ppp χ^2_{MH} .

From the results it can be concluded that, according to the fractional Bayes factors, none of the eight items appear to be measurement invariant. The support in favour of measurement non-invariance is lowest for item 6, where the FBF_{02} estimate shows just around 3.94 times (1/.254) more support for H_2 compared to H_0. Item 6, which concerns the question of whether the country's cultural life is undermined or enriched by immigrants, shows the strongest support for measurement invariance. The item with the highest degree of measurement variance appears to be item 3, where τ is estimated to be 0.149. For this item, respondents were asked their opinion with respect to the country's crime problems. For the other six items, a large degree of support was found in favour of measurement non-invariance, with τ' ranging from .036 to 0.080.

A discrepancy can be observed between the results for the fractional Bayes factors FBF_{0u} and FBF_{02} and the results for the ppp χ^2_{MH} . The latter appears to indicate that for all items there is a substantial discrepancy between the model (in which measurement invariance is assumed) and the observed data. The most noticeable difference between the result for the fractional Bayes factor and the result for the ppp χ^2_{MH} is present for item 2. The FBF_{02} indicates that it is approximately 30 times more likely that item 2 is measurement variant than not: the ppp χ^2_{MH} is just higher than .05, providing some evidence that the item might not be measurement invariant. With a strict cut-off value of .05, the conclusion would be that there is no evidence that the measurement invariance hypothesis (H_0) should be rejected, which is in contrast with the conclusion based on the results for the fractional Bayes factors.

For randomly selected groups, measurement non-invariance was assessed by using data from the countries Austria, Belgium, Czech Republic, Denmark, Germany and Switzerland. Although the estimated degree of measurement variance differed strongly between items, it was remarkable that each of the eight items showed a large degree of support for the measurement non-invariance hypothesis over the measurement invariance hypothesis. The item with the highest degree of measurement variance was again item 3,

with an estimated τ' of .184. The other seven items were considered moderately measurement variant, with estimated τ' values ranging from 0.031 to 0.080.

Table 4.4. Results for estimating the degree of measurement variance for items from the ESS

	Fixed Groups						Random Groups				
Item	τ'	ln(FBF_{0u})	FBF_{0u}	ln(FBF_{02})	FBF_{02}	ppp χ^2_{MH}	τ'	ln(FBF_{0u})	FBF_{0u}	ln(FBF_{02})	FBF_{02}
1	0.080	−18.870	<0.001	−18.870	<0.001	0.000	0.075	−123.900	<0.001	−119.325	<0.001
2	0.036	−3.569	0.028	−3.509	0.030	0.053	0.047	−79.322	<0.001	−74.747	<0.001
3	0.149	−74.763	<0.001	−74.763	<0.001	0.000	0.184	−323.815	<0.001	−319.239	<0.001
4	0.055	−6.244	0.002	−6.231	0.002	0.000	0.031	−49.707	<0.001	−45.132	<0.001
5	0.069	−8.791	<0.001	−8.791	<0.001	0.000	0.040	−63.955	<0.001	−59.379	<0.001
6	0.022	−1.772	0.170	−1.369	0.254	0.022	0.049	−82.171	<0.001	−77.596	<0.001
7	0.058	−12.931	<0.001	−12.934	<0.001	0.000	0.080	−140.381	<0.001	−135.806	<0.001
8	0.073	−20.636	<0.001	−20.636	<0.001	0.002	0.042	−67.232	<0.001	−62.656	<0.001

Notes: FBF_{01}: fractional Bayes factor, where H_0:τ =0 and H_u: $\tau \neq 0$; FBF_{02} = fractional Bayes factor, where H_0: τ=0 and H_2: τ>0; ppp χ^2_{MH}: posterior predictive p-value based on the Mantel-Haenszel χ^2_{MH} statistic.

Conclusion and Discussion

A marginal measurement invariance test has been discussed for detecting measurement non-invariance using a marginal random item effects model. This method uses the additional correlation between observations in order to detect the presence of measurement non-invariance without conditioning on group-specific item parameters. That is, one common (measurement invariant) item parameter that applies to all groups is modelled. As a result, any group-specific deviations are included in the errors. Subsequently, measurement non-invariance can be detected by evaluating the correlation between residuals within a group. The functioning of this method for the detection of measurement non-invariance was evaluated with simulation studies and applied to empirical data.

The simulation studies showed that this new method is able to estimate the degree of measurement variance for both randomly selected and fixed (stratified) groups. The fractional Bayes factor was able to accurately determine whether the estimated degree of measurement variance was equal to or greater than zero, and it outperformed the posterior predictive test based on the MH statistic. The results for the randomly selected groups were more convincing compared to the results for the fixed groups, because both within-group and between-group information was used in evaluating the level of measurement variance in the randomly selected groups.

For fixed groups when measurement invariance was assumed, the data showed support for parameter values around zero for the specified simulated conditions, which led to slightly more support for the alternative hypothesis of no measurement invariance (H_u). The fractional Bayes factor was less than one but did not show significant support for H_u. When the alternative hypothesis was specified to be measurement non-invariance (H_2), a large degree of support in favour of the measurement invariance hypothesis (H_0) was found under simulated conditions.

The posterior mean is used as a point estimator of the covariance parameter, which has a skewed posterior distribution. When measurement variance is relatively low and the distribution is right-skewed, the posterior mean tends to overestimate the degree of measurement variance. When the degree of measurement variance is relatively high and

the distribution is left-skewed, the posterior mean tends to underestimate the degree of measurement variance. This is a property of the posterior mean as a point estimator and does not relate to the properties of the proposed fractional Bayes factors, whose computations are based on sampled values from the posterior, taking into account any skewness of the posterior.

The marginal test approach has great potential and improves current methods in different ways. An overview of the advantages can be given:

- All scale items can be tested simultaneously (i.e. a sequential procedure is not needed, also dependent hypotheses can be tested simultaneously). That is: full, partial or single measurement invariance hypotheses can be tested simultaneously to avoid the risk of capitalisation on chance. The simultaneous evaluation of multiple measurement invariance hypotheses works in a similar way as testing a single measurement invariance hypothesis.
- The data evidence can be quantified in favour of partial or full measurement invariance, which is usually considered to be the null hypotheses. This is in contrast to frequentist hypothesis testing, where the null hypothesis is rejected, when significant evidence is found in favour of an alternative hypothesis.
- Uninformative and informative priors can be specified in testing measurement invariance assumptions, where the amount of prior information can be fully controlled. This makes it possible to use additional information, beside the sample data, to make inferences about measurement invariance.
- In the BF test both measurement invariance hypotheses do not need to be true to make valid inferences. The BF test only provides information about which hypothesis is more likely, and in the decision both hypotheses are taken into account. This is in contrast to frequentist hypothesis testing, where inferences are made by assuming that the null hypothesis is true.
- The marginal modelling approach makes it possible to interpret results on a common scale and they are statistically comparable without needing anchor items. Factor scores can be compared across groups even when all items are identified as measurement variant.
- The marginal test approach can be used for non-random groups, to make inferences about differences between specific groups in the sample. It can also be used for groups sampled from a population, to make inferences about the measurement invariance assumptions in the population. The complexity of the method does not increase when increasing the number of groups.
- The marginal test produces exact results and does not rely on asymptotic theory. The validity of the test results does not depend on the sample size, which makes the method also usable for small data sets.

Although the shown examples were limited to binary scored items, a balanced design, and the one-parameter IRT model, the methodology can be extended to included different data types, explanatory information, different modelling levels, multidimensionality, and so forth.

The extension to polytomous data can be established by using the marginal random item effects model with random threshold parameters as discussed by Fox (2010, pp. 193-225[37]), also referred to as the random item effects model for polytomous data.

Subsequently, the covariance structure of augmented responses to each response category needs to be examined to evaluate the measurement invariance assumptions of the threshold parameters of the items.

The marginal test approach for unbalanced data requires a numerical integration method to compute the marginal distribution of the data under each hypothesis. The integration over the covariance parameters cannot be done analytically due to the unbalanced design. However, a simulation technique, such as importance sampling, can be used to numerically evaluate the integrals required to compute the BF.

The extension to evaluate metric invariance requires a different construction of the BF. In that case, the covariance structure implied by the random item effects model contains the variation in discrimination across groups. Consider the random item effects model with random difficulty and discrimination parameters for continuous responses

$$Z_{ijk} = a_{jk}\theta - b_{jk} + \varepsilon_{ijk}$$
$$a_{jk} \sim N(a_k, \sigma_a^2) \qquad \text{Equation 4.23}$$
$$b_{jk} \sim N\left(b_k, \sigma_b^2\right)$$

The marginal model, after integrating out the random item parameters, is a multivariate normal model for the responses to item *k* in group *j* with covariance matrix

$$\boldsymbol{\Sigma}_{jk} = \sigma_a \boldsymbol{\theta}_j \boldsymbol{\theta}_j^t + \sigma_b \mathbf{J}_m + \mathbf{I}_m \qquad \text{Equation 4.24}$$

It can be seen that the first term, $\sigma_a \boldsymbol{\theta}_j \boldsymbol{\theta}_j^t$, represents the dependency caused by a random discrimination effect, and the second term $\sigma_b \boldsymbol{J_m}$ by a random difficulty effect. For a measurement invariant item, the variances σ_a and σ_b are zero and the responses are independently distributed. A positive correlation implies additional dependencies between the responses, which represents a violation of measurement invariance. Bayes factor tests can be defined to test hypotheses about the variance parameters σ_a and σ_b. When $\sigma_a = 0$ the discussed BFs can be used to evaluate measurement invariance hypotheses about the difficulty parameter.

There are two ways to include survey weights into the analysis. The most straightforward approach is to weight the likelihood, where the weights function as frequency weights (Rabe-Hesketh and Skrondal, 2006[61]). A pseudo-likelihood can be constructed, which is used to construct the posterior distributions. A disadvantage is that the weights should be partitioned to identify the level-specific weights. Most often the general inclusion probabilities are given and not the level-specific inclusion probabilities. The construction of a pseudo-likelihood also leads to a computational complexity, since samples cannot be directly drawn from the posterior distribution. A Metropolis-Hastings algorithm could be used to draw samples from the posterior distribution to facilitate estimation of the model parameters and to compute the fractional Bayes factor. However, more research is needed to evaluate the strengths of a pseudo-likelihood approach.

Another approach is to weigh the (underlying) latent response data to reflect the unequal sampling probabilities. In that case, the weights can address additional correlations in the data that are not explicitly modelled. The advantage of weighing the latent responses is that

the computational approach remains the same as well as the construction of the Bayes factors. Future research is focused on the modelling of weighted latent responses to deal with complex relationships in the data, while making use of the computational algorithms for the non-weighted responses.

This chapter shows that the model can be applied to empirical data. Data regarding the attitude towards immigration questions in the ESS was used to illustrate the method. The results show that measurement non-invariance appears to be present in all items included in this empirical example. The marginal modelling approach accommodates the measurement non-invariance in the difficulty parameters by modelling the implied correlations between the responses. Therefore, computed factor scores under the marginal model are comparable across groups. This makes the marginal model also suitable for computing comparable factor scores, when all items exhibit measurement non-invariance.

Annex 4.A. Specification of Priors and Posterior Distributions

An orthogonal matrix is used to transform the latent response data in two components, where one component represents the information concerning, τ_k, and the other component the information concerning the measurement error variance. A Helmert matrix $\boldsymbol{H}$ (Lancaster, 1965[62]) of dimension m by m is used to transform the latent responses. The first row has elements $1/\sqrt{m}$, and the remaining rows below are $\left[\frac{\mathbf{1}_s^t}{\sqrt{s(s+1)}}, \frac{-1}{\sqrt{s(s+1)}}, \mathbf{0}_{m-s-1}\right]$ for $s = 2, \ldots, m$. Consider the transformation $\tilde{\mathbf{z}}_{jk} = \mathbf{H}\mathbf{z}_{jk}$. When assuming the covariance structure in Equation 4.17, it can be shown that the first component is normally distributed with mean and variance equal to

$$E\left(\tilde{z}_{jk1}\right) = E\left(\sqrt{m}\bar{z}_{jk}\right) = \sqrt{m}\left(\bar{\theta}_j - b_k\right)$$
$$Var\left(\tilde{z}_{jk1}\right) = \left[m\left(1+\tau_k\right) + m\left(m-1\right)\tau_k\right] / m = 1 + m\tau_k, \qquad \textbf{Equation 4.25}$$

respectively, and the remaining components are normally distributed with mean zero and variance one. Note that the transformed variables are independently distributed, since it is an orthogonal transformation.

To make inferences about the covariance parameter, consider the distribution of the first transformed variable. It follows that

$$\begin{aligned} p\left(\tilde{\mathbf{z}}_{k1} | \boldsymbol{\theta}, b_k, \tau_k\right) &\propto \left(1+m\tau_k\right)^{-J/2} \exp\left(\frac{-mS_B}{2\left(1+m\tau_k\right)}\right) \\ &\propto \left(1/m+\tau_k\right)^{-J/2} \exp\left(\frac{-S_B}{2\left(1/m+\tau_k\right)}\right) \end{aligned} \qquad \textbf{Equation 4.26}$$

where

$$mS_B = \sum_{j=1}^{J}\left(\tilde{z}_{jk1} - \sqrt{m}\left(\bar{\theta}_j - b_k\right)\right)^2 = m\sum_{j=1}^{J}\left(\bar{z}_{jk1} - \left(\bar{\theta}_j - b_k\right)\right)^2 \qquad \textbf{Equation 4.27}$$

A non-informative reference prior is defined for τ_k, which is given by

$$p\left(\tau_k\right) \propto \left(1/m+\tau_k\right)^{-1} \qquad \textbf{Equation 4.28}$$

Then, the posterior distribution of the covariance parameter is given by

$$p\left(\tau_k \left| \boldsymbol{\theta}, b_k, \tau_k, \tilde{\mathbf{z}}_{k1}\right.\right) = \frac{\left(S_B\right)^{J/2}}{\Gamma\left(\frac{J}{2}\right)}\left(1/m + \tau_k\right)^{-J/2-1} \exp\left(\frac{-S_B}{2\left(1/m + \tau_k\right)}\right) \quad \textbf{Equation 4.29}$$

which can be recognised as a shifted-inverse gamma distribution, with shape parameter *J/2*, rate parameter $S_B/2$, and shift parameter $1/m$ [see also Fox, Mulder, and Sinharay, (2017[42])]. The parameter τ_k is sampled from the posterior distribution by sampling $\lambda_k = 1/m + \tau_k$ from the inverse-gamma distribution with shape parameter *J/2* and scale parameter $S_B/2$, to obtain a draw $\tau_k = \lambda_k - 1/m$.

A similar procedure is applied for the fixed group situation, where the covariance matrix of the latent response data is given in Equation 4.18. For this covariance structure, the transformed latent response data are normally distributed with the mean and variance of the first transformed component equal to

$$\begin{aligned} E\left(\tilde{z}_1\right) &= E\left(\sqrt{m}\bar{z}_{jk}\right) = \sqrt{m}\left(\bar{\theta}_j - b_k\right) \\ Var\left(\tilde{z}_1\right) &= 1 + \left(m-1\right)\tau_k \end{aligned} \quad \textbf{Equation 4.30}$$

and the remaining components are normally distributed with mean zero and variance $1-\tau_k$. Then, the distribution of the first transformed component is given by

$$p\left(\tilde{\mathbf{z}}_{k1} \left| \boldsymbol{\theta}, b_k, \tau_k\right.\right) \propto \left(1/\left(m-1\right) + \tau_k\right)^{-J/2} \exp\left(\frac{-mS_B/(m-1)}{2\left(1/\left(m-1\right) + \tau_k\right)}\right) \quad \textbf{Equation 4.31}$$

It follows that $\tau_k \geq -1/(m-1)$, and a non-informative prior for τ_k is specified, $p\left(\tau_k\right) \propto \left(1/(m-1) + \tau_k\right)^{-1}$, which leads to a shifted-inverse gamma posterior distribution for τ_k given the first transformed component:

$$p\left(\tau_k \left| \boldsymbol{\theta}, b_k, \tau_k, \tilde{\mathbf{z}}_{k1}\right.\right) = \frac{\left(\frac{\tilde{S}_B}{2}\right)^{-J/2}}{\Gamma\left(\frac{J}{2}\right)}\left(1/\left(m-1\right) + \tau_k\right)^{-J/2-1} \exp\left(\frac{-\tilde{S}_B/2}{1/\left(m-1\right) + \tau_k}\right) \quad \textbf{Equation 4.32}$$

where $\tilde{S}_B = mS_B/(m-1)$.

However, in this case the remaining transformed components also contain information about τ_k. The distribution of the remaining transformed components is given by

$$p\left(\tilde{\mathbf{z}}_{2k},\ldots,\tilde{\mathbf{z}}_{mk}\,|\boldsymbol{\theta},b_k,\tau_k\right)\propto\left(\sigma_{\varepsilon k}^2\right)^{-J(m-1)/2}\exp\left(\frac{-S_w}{2\sigma_{\varepsilon k}^2}\right), \qquad \textbf{Equation 4.33}$$

where $\sigma_{\varepsilon k}^2=1-\tau_k$ and $S_w=\sum_{j=1}^{J}\sum_{i=2}^{m}\left(\tilde{z}_{ijk}\right)^2$. Through a variable transformation, the non-informative prior for $\sigma_{\varepsilon k}^2$ equals $p\left(\sigma_{\varepsilon k}^2\right)\propto\sigma_{\varepsilon k}^{-2}$. This leads to an inverse-gamma posterior distribution for the $\sigma_{\varepsilon k}^2=1-\tau_k$ given the transformed variables $\tilde{\mathbf{z}}_{ik},i=2,\ldots,m$, with shape parameter *J(m-1)* and rate parameter $mS_B/(m-1)$:

$$p\left(\sigma_{\varepsilon k}^2\,|\boldsymbol{\theta},b_k,\tilde{\mathbf{z}}_{2k},\ldots,\tilde{\mathbf{z}}_{mk}\right)=\frac{\left(\frac{S_w}{2}\right)^{-J(m-1)/2}}{\Gamma\left(\frac{J(m-1)}{2}\right)}\left(\sigma_{\varepsilon k}^2\right)^{-J(m-1)/2}\exp\left(\frac{-S_w}{2\sigma_{\varepsilon k}^2}\right). \qquad \textbf{Equation 4.34}$$

To sample values for τ_k, the $\tilde{\tau}_k$ is sampled from the shifted-inverse gamma distribution and $\sigma_{\varepsilon k}^2=1-\tau_k$ from the inverse-gamma distribution (see the MCMC algorithm) (Fox, Mulder and Sinharay, 2017[42]). Then, the intra-class correlation, $\frac{\tilde{\tau}_k}{\sigma_{\varepsilon k}^2+\tilde{\tau}_k}=\tau_k$, is constructed from the sampled values of $\tilde{\tau}_k$ and $\sigma_{\varepsilon k}^2$ to obtain a final draw for the covariance parameter τ_k. When sampling the $\sigma_{\varepsilon k}^2=1-\tau_k$ and $\tilde{\tau}_k$ in different steps, we found that the statistical inferences on the sampled values are complicated, since the sum of both components is restricted to one. A more efficient inference about τ_k can be made when the posterior information about $\sigma_{\varepsilon k}^2=1-\tau_k$ is included. Therefore, the intraclass-correlation coefficient is considered, which is given by $\tau_k/(\tau_k+\sigma_{\varepsilon k}^2)$. For the covariance structure defined in Equation 4.18, it is equal to the correlation coefficient τ_k. The intraclass-correlation is not scale dependent, which makes it possible to sequentially sample a value for τ_k and $\sigma_{\varepsilon k}^2$, without the restriction that the sampled values sum to one. As a result, each computed intraclass-correlation given the sampled parameter values is a sampled value of parameter τ_k, and given the sampled values, posterior inferences can be made about τ_k.

Annex 4.B. Fractional Bayes Factor

The object is to define the fractional Bayes factor for the hypotheses $H_0 : \tau_k = 0$, $H_2 : \tau_k > 0$ and $H_u : \tau_k \neq 0$ for random and fixed group situation. First consider the covariance structure defined in Equation 4.17, which represents the random group situation. Assume a total of *N* responses to item *k*, and a balanced design for *J* groups with each *m* group members. The marginal distribution of the data under H_0 is given by

$$\begin{aligned} p(\mathbf{z}; s = 1/J, H_0) &= \frac{p(\tilde{\mathbf{z}}_1 | \tau_k, \mathbf{b}, H_0)\, p(\tilde{\mathbf{z}}_2, \ldots, \tilde{\mathbf{z}}_m | \mathbf{b}, H_0)}{p(\tilde{\mathbf{z}}_1 | \tau_k, \mathbf{b}, H_0)^{1/J}} \\ &= \frac{(2\pi)^{\frac{-Jm}{2}} \exp\left(-\frac{1}{2}(S_w + mS_B)\right)}{(2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(mS_B / 2J)\right)} \\ &= (2\pi)^{-\frac{(Jm-1)}{2}} \exp\left(-\frac{1}{2}(S_w + mS_B(1 - 1/J))\right), \end{aligned} \qquad \textbf{Equation 4.35}$$

where $S_w = \sum_{j=1}^{J}\sum_{i=2}^{m}(\tilde{z}_{ijk})^2$ and $S_B = \sum_{j=1}^{J}(\bar{z}_{jk1} - (\bar{\theta}_j - b_k))^2$.

The marginal distribution of the data under H_u requires integration over the parameter space of the covariance parameter. It follows that

$$\begin{aligned} p(\mathbf{z}; s = 1/J, H_u) &= \frac{\int p(\tilde{\mathbf{z}}_1 | \tau_k, \mathbf{b}, H_u)\, p(\tilde{\mathbf{z}}_2, \ldots, \tilde{\mathbf{z}}_m | \mathbf{b}, H_u)\, p(\tau_k)\, d\tau_k}{\int p(\tilde{\mathbf{z}}_1 | \tau_k, \mathbf{b}, H_0)^{1/J}\, p(\tau_k)\, d\tau_k} \\ &= \frac{p(\tilde{\mathbf{z}}_2, \ldots, \tilde{\mathbf{z}}_m | \mathbf{b}, H_u) \int p(\tilde{\mathbf{z}}_1 | \tau_k, \mathbf{b}, H_u)\, p(\tau_k)\, d\tau_k}{\int p(\tilde{\mathbf{z}}_1 | \tau_k, \mathbf{b}, H_0)^{1/J}\, p(\tau_k)\, d\tau_k} \\ &= (2\pi)^{-\frac{(Jm-1)}{2}} \exp(-S_w / 2) \frac{\Gamma\left(\frac{J}{2}\right)\left(\frac{mS_B}{2}\right)^{-J/2}}{\Gamma\left(\frac{1}{2}\right)\left(\frac{mS_B}{2J}\right)^{-1/2}}, \end{aligned} \qquad \textbf{Equation 4.36}$$

where the integration over τ_k is performed using the fact that $\tau_k + m^{-1}$ has an inverse-gamma distribution. In the same way, the marginal distribution of the data under H_2 is derived as

$$
\begin{aligned}
p\left(\mathbf{z}; s = 1/J, H_2\right) &= \frac{p\left(\tilde{\mathbf{z}}_2,\ldots,\tilde{\mathbf{z}}_m \middle| \mathbf{b}, H_u\right)\int_0^\infty p\left(\tilde{\mathbf{z}}_1 \middle| \tau_k, \mathbf{b}, H_u\right) p\left(\tau_k\right) d\tau_k}{\int_0^\infty p\left(\tilde{\mathbf{z}}_1 \middle| \tau_k, \mathbf{b}, H_0\right)^{1/J} p\left(\tau_k\right) d\tau_k} \\
&= (2\pi)^{-\frac{-(Jm-1)}{2}} \exp\left(-S_w/2\right) \frac{\Gamma\left(\frac{J}{2}\right)\left(\frac{mS_B}{2}\right)^{-J/2}}{\Gamma\left(\frac{1}{2}\right)\left(\frac{mS_B}{2J}\right)^{-1/2}} \frac{1 - F\left(J/2, mS_B/2, 1/m\right)}{1 - F\left(1/2, mS_B/2J, 1/m\right)},
\end{aligned}
\qquad \textbf{Equation 4.37}
$$

where $F\left(\alpha, \beta, \gamma\right)$ denotes the cumulative shifted-inverse gamma distribution with shape parameter α, rate parameter β and shift parameter γ. The ratio of the defined marginal distributions can be used to construct a fractional Bayes factor, and to test hypothesis concerning the covariance parameter.

Each fractional Bayes factor is computed in each MCMC iteration, given the person and item parameters, and the latent response data $\mathbf{z}_k$. The mean estimate across MCMC iterations is an estimate of the final fractional Bayes factor. The marginal distribution of the data also depends on the item difficulty and person parameters, but they are only involved in the between-group, S_B, and within-group sum of squares, S_w, of the latent response data. The integration over the item and person parameters is facilitated through the MCMC algorithm.

Next, consider the covariance structure defined in Equation 4.18, which represents the fixed group situation. For this covariance structure, improper priors are used for parameters τ_k and $\sigma^2_{\varepsilon k}$. A minimum informative sample is used in the fractional Bayes factor, where $s_1 = N_1^{-1} = 1/(J(m-1))$ and $s_2 = 1/J$, to deal with the improper priors. Under the null hypothesis, representing measurement invariance for item k, τ_k =0 and $\sigma^2_{\varepsilon k}$ =1. Then, the marginal distribution of the data under the null hypothesis equals

$$
\begin{aligned}
p\left(\mathbf{z}_k, s_1 = N_1, s_2 = 1/J, H_0\right) &= \frac{p(\tilde{\mathbf{z}}_{2k},\ldots,\tilde{\mathbf{z}}_{mk} \mid \sigma^2_{\varepsilon k} = 1)\, p(\tilde{\mathbf{z}}_{1k} \mid \tau_k = 0)}{p(\tilde{\mathbf{z}}_{2k},\ldots,\tilde{\mathbf{z}}_{mk} \mid \sigma^2_{\varepsilon k} = 1)^{1/N_1}\, p(\tilde{\mathbf{z}}_{1k} \mid \tau_k = 0)^{1/J}} \\
&= (2\pi)^{-(N/2-1)} \exp\left(-1/2\left(S_w(1 - N_1^{-1}) + S_b(1 - J^{-1})\right)\right),
\end{aligned}
\qquad \textbf{Equation 4.38}
$$

where $S_b = \frac{m}{m-1} S_b$. For the unrestricted hypothesis, the marginal distribution of the latent response data to item *k* is obtained by integrating out the covariance parameters in the expressions for the Helmert-transformed data. Under the unrestricted hypothesis, parameter $\sigma^2_{\varepsilon k}$ is defined on the interval $(0,1)$, referred to as H_{u_σ}; parameter τ_k is defined on $\left[-1/(m-1),\ 1\right]$, referred to as H_{u_τ}. It follows that

$$p(\mathbf{z}_k, s_1, s_2, H_u) = \frac{\int_{\sigma^2_{\varepsilon k} \in H_{u_\sigma}} p(\tilde{\mathbf{z}}_k \mid \sigma^2_{\varepsilon k}) p(\sigma^2_{\varepsilon k}) d\sigma^2_{\varepsilon k} \int_{\tau_k \in H_{u_\tau}} p(\tilde{\mathbf{z}}_k \mid \tau_k) p(\tau_k) d\tau_k}{\int_{\sigma^2_{\varepsilon k} \in H_{u_\sigma}} p(\tilde{\mathbf{z}}_k \mid \sigma^2_{\varepsilon k})^{1/N_1} p(\sigma^2_{\varepsilon k}) d\sigma^2_{\varepsilon k} \int_{\tau_k \in H_{u_\tau}} p(\tilde{\mathbf{z}}_k \mid \tau_k)^{1/J} p(\tau_k) d\tau_k}$$

$$= (2\pi)^{-(N/2-1)} \frac{\Gamma\left(\frac{N_1}{2}\right)\Gamma\left(\frac{J}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^2} \left(\frac{S_w}{2}\right)^{-\frac{N_1}{2}} \left(\frac{\tilde{S}_b}{2}\right)^{-\frac{J}{2}} \left(\frac{S_w}{2N_1}\right)^{\frac{1}{2}} \left(\frac{\tilde{S}_b}{2J}\right)^{\frac{1}{2}} \times \quad \textbf{Equation 4.39}$$

$$\frac{F\left(1; N_1/2, S_w/2, 0\right)}{F\left(1; 1/2, S_w/2N_1, 0\right)} \frac{F\left(1; J/2, \tilde{S}_b/2, 0\right)}{F\left(1; 1/2, \tilde{S}_b/2J, 0\right)},$$

where the last two terms follow from the truncation of the parameters to the intervals H_{u_σ} and H_{u_τ}. For hypothesis $H_2; \tau_k > 0,$ the marginal distribution of the data in Equation 4.39 is slightly modified, since the integration of τ_k is restricted to (0,1). This leads to a small modification of the cumulative inverse-gamma probability concerning parameter τ_k, and the last term on the right-hand side of Equation 4.39 becomes

$$\frac{F\left(1, \frac{J}{2}, \frac{S_b}{2}, 0\right) - F\left(\frac{1}{m-1}; \frac{J}{2}, \frac{S_b}{2}, 0\right)}{F\left(1; \frac{1}{2}, \frac{S_b}{2J}, 0\right) - F\left(\frac{1}{m-1}; \frac{1}{2}, \frac{S_b}{2J}, 0\right)}. \quad \textbf{Equation 4.40}$$

Annex 4.C. Simulation Study

A simulation study was conducted using our own programme developed in R (R Core Team, 2014[63]). In this first simulation study, binary response data were simulated for 10 items, with 1 000 persons assigned to one of two groups. The degree of measurement variance τ_k was increased across items. The lower bound of $\tau_k = -1/m$, where $m = 500$ represents the number of persons per group. For item 1, the level of measurement variance equaled this lower bound. The simulation study consisted of 50 data replications, which provided stable results; the mean results across replications are reported.

The MCMC algorithm was used to estimate the degree of measurement variance in each item. The number of MCMC iterations was set to 5 000 with a burn-in of 1 000. The convergence and autocorrelation plots, created using the R package (Plummer et al., 2006[64]) showed no irregularities. The functioning of the fractional Bayes factor was compared to the functioning of the ppp χ^2_{MH} statistic for the detection of measurement non-invariance.

Sinharay, Johnson and Stern (2006[65]) showed that the χ^2_{MH} statistic is useful in assessing model fit in posterior predictive model checking. They used the χ^2_{MH} statistic in order to test for local independence, where responses to items are assumed to be independently distributed given the person parameter. The association among item pairs was investigated to detect possible violations of the local independence assumption. This relates to the assumption of measurement invariance, where responses to item k are assumed to be independently distributed given a common item difficulty parameter for the reference and focal groups. Therefore, it is to be expected that the statistic can also be used to test measurement invariance assumptions. When responses to item k are independently distributed given the item parameter and group membership of the respondents (i.e. reference or focal group), it is concluded that measurement invariance does not hold. Data are replicated under the model, where it is assumed that the degree of measurement variance $\tau = 0$. The posterior predictive p-value (ppp χ^2_{MH}) is estimated by the proportion of MCMC iterations in which the value of the χ^2_{MH} statistic for the replicated data is greater than the one for the observed data:

$$P\left(\chi^2_{\text{MH}}\left(\mathbf{y}_{\text{rep},}\right) \geq \chi^2_{\text{MH}}\left(\mathbf{y}_{obs}\right) | \mathbf{y}_{obs}\right) \qquad \textbf{Equation 4.41}$$

This simulation study involved 50 data replications, and the mean of the ppp χ^2_{MH} over 50 replications was computed. The estimated ppp χ^2_{MH} represents the extremeness of the statistic for the observed data using replicated data generated under the assumption of measurement invariance. When the observed statistic value was extreme under the assumption of measurement invariance, a violation of this assumption was detected. A ppp

χ^2_{MH} of .5 indicates that the measurement invariance assumption is not violated, whereas a value close to 0 indicates that it is (Sinharay, Johnson and Stern, 2006[65]). However, as Gelman, Meng and Stern (1996[66]) pointed out, the ppp χ^2_{MH} shows the degree to which there are discrepancies between the model and the observed data. They emphasise that it is more of a tool to assess the usefulness of a model than a test to determine whether or not the model is true.

In the second simulation study, a dataset was generated with 1 000 persons, equally divided over 20 randomly selected groups. The responses (either incorrect or correct) of these 1 000 persons were simulated over 10 items. The degree of measurement variance τ increased across items, as it did in the first simulation study. The lower bound was $-1/m.$ Here, the lowest possible value for measurement non-invariance would be $-1/50.$. The fractional Bayes factors were computed to detect evidence in favour of the measurement invariance hypothesis H_0, when the alternative hypotheses are no measurement invariance H_u and measurement non-invariance H_2.

The number of MCMC iterations was 5 000 with a burn-in of 1 000. The convergence and autocorrelation plots, created using the R package coda (Plummer et al., 2006[64]) did not show any irregularities. As in the previous study, this study consisted of 50 data replications, which led to stable results.

Chapter 5. Multigroup and Multilevel Latent Class Analysis

Michael Eid

Introduction

Latent class analysis (LCA) is a statistical model for explaining the associations between observed categorical variables by the existence of latent categorical variables (Clogg, 1995[67]; Collins and Lanza, 2010[68]; Hagenaars and McCutcheon, 2002[69]; Lazarsfeld and Henry, 1968[70]). The starting point of an LCA are observed categorical variables that are measured on a nominal or an ordinal scale. Bartholomew, Knott and Moustaki (2011[71]) distinguish four types of latent variable models depending on the nature of the observed and the latent variables:

1. factor analytic models are models for continuous observed and continuous latent variables
2. latent trait models are models for observed categorical and continuous latent variables
3. latent class models are models for categorical observed and categorical latent variables
4. latent profile models are models for continuous observed and categorical latent variables.

Hence, latent class models are applied if items have been assessed by a categorical response format and a researcher does not assume that the items can be ordered on one or more latent continuous variables (such as in a latent trait model). Instead, it is assumed that there are latent typological differences represented by latent classes. The latent classes are the values of a nominal-scaled latent variable. The categories of a nominal scale represent qualitative differences. However, the latent classes can be ordered as a result of the analysis (Heinen, 1996[72]). This makes it possible to use LCA to prove whether different items can be ordered on latent dimensions and whether a dimensional model is reasonable in an application. Hence, LCA is a much more general approach compared to other models with latent variables such as unidimensional models of item response theory.

LCA is a very general approach for considering measurement error and for reducing the number of observed response patterns to a smaller number of latent classes. Measurement error is considered by the fact that the membership of a latent class does not determine the observed response perfectly but only with a certain probability. Hence, the link between an observed and a latent categorical variable is given by the conditional response probabilities for the categories of the observed variables given a category of the latent class variable. Latent class models can also be considered as multinomial logit models with multiple categorical dependent variables and a latent categorical independent variable.

Multigroup LCA allows to compare different groups (e.g. countries, genders) with respect to a latent class model. For example, it can be analysed whether the same latent classes can be found in different groups and whether the classes have the same sizes. If the groups are randomly selected from a population of groups (e.g. schools in a country) *multilevel LCA* can be applied, for example, to compare the variation of class sizes across groups.

Latent class models have been applied in quite different areas of research (Hagenaars and McCutcheon, 2002[69]) and have become popular in recent studies in the context of educational research and large-scale assessment. Drossel and Eickelmann (2017[73]), for example, used LCA to cluster German and Czech teachers into subgroups differing in types of activities that are related to technology-related professional development. Zhang, Watermann and Daniel (2016[74]) found different latent achievement goals classes and analysed their associations with achievement test scores. Based on LCA, Oliveri et al. (2014[75]) analysed differential item functioning of items of the Progress in International Reading Literacy Study (PIRLS) that were administered in Hong Kong, Chinese Taipei, Kuwait and Qatar. Fagginger Auer et al. (2016[76]) applied multilevel LCA to scrutinise the relationship between the curriculum and different mathematical strategies used by students. Yalcin (2017[77]) used data from the Turkish PISA study to analyse the determinants of different achievement classes by multilevel LCA. Using the PISA data for Chinese Taipei, Lin and Tai (2015[78]) analysed in which way different latent classes of mathematics learning strategies are related to the mathematical literacy of students. Boyce and Bowers (2016[79]) detected different latent classes of principals who quit their schools. These few examples show that LCA can be fruitfully applied in educational and large-scale studies.

Description of LCA and its Extensions to Multigroup and Multilevel Models

Basic Assumptions

Latent Class Analysis

In classical LCA it is assumed that the population consists of a finite set of sub-populations (so-called latent classes) that are disjoint and exhaustive. Disjoint means that the latent classes do not overlap and a member of the total population belongs only to one latent class. Exhaustive means that each member of the total population has to belong to a latent class. It is not known in advance to which latent class a member of the population belongs. After an LCA has been conducted the probabilities to belong to the different classes can be estimated for each member of the population based on his or her response pattern (assignment probabilities). Based on the assignment probabilities an individual can be assigned to the latent class for which his or her assignment probability is maximal. The latent classes can differ in their sizes. The probability that a randomly selected member belongs to a latent class (latent class probabilities) can be estimated.

Each latent class is characterised by the response probabilities for the categories of the observed variables. These conditional response probabilities are the same for all members of a latent class, the different latent classes differ at least in the response probabilities of one item (observed variable). The latent classes are defined by these response probabilities and their substantive meaning can be delineated by interpreting the patterns of the class-specific (conditional) response probabilities. Finally, it is assumed that all observed items are stochastically independent given the latent classes (so-called local independence). Local independence means that within a single latent class the observed variables are independent. The latent classes explain the unconditional associations of the observed variables.

Multigroup Latent Class Analysis

Multigroup LCA is an extension of LCA by considering multiple groups (Clogg and Goodman, 1985[80]; Eid, Langeheine and Diener, 2003[81]; Kankaraš, Moors and Vermunt, 2018[82]). The different groups have to be independent from each other and the membership

to the different groups has to be known. In OECD data sets the different groups are, for example, different countries. Multigroup LCA allows testing specific hypotheses about the latent class structure in different countries. For example, a researcher conducting a multi-country study might be interested in different research questions:

- Do the countries differ in the number of latent classes?
- Do the latent classes have the same meaning in the different countries?
- Do the latent classes have the same sizes in the different countries?

We illustrate how multigroup LCA can give answers to these research questions by referring to the distributed leadership scale of the TALIS 2013 database measuring participation in school decisions assessed by the principal of the school (OECD, 2014[83]). This scale consists of the following five statements (TC2G22A to TC2G22E) that have to be answered on a rating scale with four categories (1-strongly disagree, 2-disagree, 3-agree, 4-strongly agree):

a This school provides staff with opportunities to actively participate in school decisions.

b This school provides parents or guardians with opportunities to actively participate in school decisions.

c This school provides students with opportunities to actively participate in school decisions.

d I make the important decisions on my own.

e There is a collaborative school culture which is characterised by mutual support.

Applying LCA to these items the countries can be compared with respect to the number of latent classes (participation types), the response probabilities and the class sizes. If the countries do not differ in the number of classes and the response probabilities, measurement invariance across countries would be given (Eid, Langeheine and Diener, 2003[81]; Kankaraš, Moors and Vermunt, 2018[82]). However, the countries are allowed to differ in their class sizes. In the context of latent class analysis, measurement invariance of a latent class is defined by showing the same conditional response probabilities across countries. If there is a latent class with the same conditional response probabilities in all countries, the meaning of this class with respect to the participation culture is the same and the sizes of this class can be compared across countries. For example, if there is a latent class in all countries showing (a) a response probability of .85 for the category *strongly agree* of the fourth item and 0.05 for the other categories of this item and (b) a response probability of .85 for the category *strongly disagree* of all other items and .05 for the other categories of all other items, this class would characterise schools in which only the principal decides and no other groups are allowed to participate in decisions. The sizes of this class in different countries can be compared to figure out whether countries differ in this type of participation culture. Full measurement invariance is given, if the countries do not differ in the number of classes and if the conditional response probabilities do not differ between countries. It is important to note that the number of classes can differ between countries even in the case of measurement invariance. For example, in Country A there can be two classes and in Country B three classes. This situation can be modelled by assuming a model with three classes and measurement invariance in both countries and fixing the size of one class in Country to 0.

Multilevel Latent Class Analysis

Multilevel LCA is an extension of LCA for analysing data from a nested sampling design, in which random samples were drawn on different levels (Vermunt, 2003[84]; 2008[85]). With respect to OECD data a two-level design is given, for example, when within each country schools and within schools teachers or students were randomly drawn. Although countries in international studies are usually not randomly chosen but selected for theoretical reasons – and the country factor is typically a fixed factor and not a random factor – it has some advantages to treat the countries as a random factor. According to Vermunt (2003[84]) a multilevel analysis with countries as a level-2 variable has the advantages that (a) the number of parameters to be estimated is reduced and (b) the estimates are usually more stable. Moreover, it makes it possible to include level-2 predictor variables to predict class membership.

Computer programmes for multilevel latent class analysis, such as M*plus* (Muthén and Muthén, 1998-2017[25]), Latent GOLD (Vermunt and Magidson, 2016[86]) and mdltm (von Davier, 2005[87]; von Davier, 2010[88]; von Davier and Rost, 2016[89]), can be used to analyse two-level LCA models. Vermunt and Magidson (2016[86]) describe how it is possible to define a three-level model making use of some options for longitudinal data analysis. However, this is a bit tricky and not standard input language. Therefore, we will focus on two-level LCA to avoid misspecifications of models. In the case that teachers within different schools within different countries will be considered, teachers are level-1 units, schools are level-2 units and countries are units of a fixed factor.

Conducting a Latent Class Analysis

Latent Class Analysis

There are in general two approaches for conducting an LCA, a confirmatory and an exploratory approach.

Confirmatory Approach

According to the confirmatory approach, specific hypotheses about the latent class structure can be tested. For example, a researcher can have the hypothesis that there are three latent classes explaining the associations between the observed variables. In order to test this hypothesis, the researcher can run an LCA using a computer programme and ask for a three-class solution. The computer programme will estimate the conditional response probabilities for the three classes and the class sizes. Based on the estimated parameters the expected frequencies for the different response vectors can be estimated. These expected frequencies can be compared with the observed frequencies with a statistical test such as the Pearson test or the likelihood ratio test (Eid, Langeheine and Diener, 2003[81]). If the test statistics show that the observed and expected frequencies do not differ significantly, the researcher keeps this model as an appropriate model for explaining the associations of the observed variables. Many more specific hypotheses about response probabilities and class sizes can be tested in a confirmatory way (Langeheine, 1988[90]). The application of the Pearson test and the likelihood ratio test, however, require large sample sizes. The expected frequency of each possible response pattern should be at least 1 or even 5 to make sure that both statistics follow a χ^2-distribution and that p values can be used for a valid decision about the model fit. If the values of the Pearson test and the likelihood ratio test differ strongly this is a sign that both do not follow a χ^2-distribution. In the case of sparse

tables bootstrapping goodness of fit measures can be applied (Langeheine, Pannekoek and Van De Pol, 1996[91]).

Exploratory Approach

However, often researchers do not have specific hypotheses, but they want to know how many classes are necessary, how large the classes are and what the meaning of the classes is. In this case, the researcher can use the LCA in an exploratory way. In an exploratory analysis one has to figure out how many classes are necessary to explain the associations of the observed variables. In order to decide about the number of classes, several latent class models with an increasing number of latent classes have to be computed. The best fitting model has to be selected. One way to compare the fit of the model is to use the likelihood ratio test and to test whether including a further latent class results in a significantly better fit. The best fitting model according to this criterion is the model for which increasing the number of latent classes would not result in a model that fits the data significantly better. However, the traditional likelihood ratio test cannot be applied because regularity conditions are violated and the bootstrap likelihood ratio test has to be used (Tekle, Gudicha and Vermunt, 2016[92]). The fit of two models can also be compared by information criteria like the Akaike Information Criterion (AIC), the Bayes Information Criterion (BIC) or the Consistent Akaike Information Criterion (CAIC). These information criteria are calculated based on the fit of the model and the number of parameters to be estimated (Yang and Yang, 2007[93]). The best fitting model is the model showing the lowest value of an information criterion. After having decided about the number of classes the conditional response probabilities and the class sizes can be interpreted. Kankaraš et al. (2018[82]) recommend the BIC and CAIC for large sample sizes, and the AIC for smaller to medium-sized sample sizes. According to simulation studies, the AIC-3 seems to perform better than the AIC (Andrews and Currim, 2003[94]; Dias, 2006[95]; Fonseca and Cardoso, 2007[96]).

Multigroup Latent Class Analysis

If there are clear hypotheses about the number and structure of the latent classes in the different groups, a confirmatory LCA can be conducted to test whether the hypothesised structure can be confirmed. However, the more typical case is that there are no clear hypotheses about the number and meaning of the latent classes. Therefore, a more exploratory analysis has to be pursued. How can this be done? There are at least two general strategies for an exploratory multigroup LCA.

Strategy I

In the first strategy, an LCA is conducted in several steps:

1. An LCA is conducted in each group (country) separately according to the exploratory approach described above to figure out how many classes have to be considered and what the conditional response probabilities tell about the meaning of the latent classes.
2. If the number of the classes is the same in all groups (countries), the assumption of full measurement invariance is tested by comparing the fit of the model without measurement invariance and the model with measurement invariance by a likelihood ratio test or – in the case of sparse tables – by a bootstrap likelihood ratio test [called "conditional bootstrap" in the computer programme Latent Gold (Vermunt and Magidson, 2005[97])] and/or by information criteria. If the assumption

of full measurement invariance has to be rejected, then different forms of partial measurement invariance can be tested. For example, it can be scrutinised whether some latent classes are measurement invariant or not and/or whether some items are measurement invariant or not. That means that the assumption of measurement invariance can be relaxed for some classes and/or items. This can be done successively until one finds such a less restrictive model that does not fit the data worse than the totally unrestricted model.

3. If the number of the classes differ between groups, then it can be tested whether the classes that are present in all groups (countries) are measurement invariant or not. For example, if there are two classes in Group A, three classes in Group B, and four classes in Group C, it can be tested whether two classes are measurement invariant across all three groups, and whether there is an additional invariant class for Group B and Group C. This model would be equivalent to a multigroup model with four latent classes in all groups and full measurement invariance. The sizes of two classes in Group A, and one class in Group B, would be 0. If this model has to be rejected, different forms of partial measurement invariance (described in the last paragraph) can be tested.

Strategy I is explained and illustrated in detail, for example, by Eid and Diener (2001[98]), Eid, Langeheine and Diener (2003[81]) as well as Kankaraš, Moors and Vermunt (2018[82]). Therefore, we will not illustrate this strategy. Strategy I can easily be applied if there is a small number of groups and items. However, this strategy has its limits if there are many countries like in the OECD studies. If there are many countries and full measurement invariance is not given across countries, it is cumbersome to find countries that consist of classes that are (partially) measurement invariant, because there a no specific fit indices to find items or classes that do not follow the measurement invariance assumption. Hence one has to compare each single parameter estimate between the restricted and the non-restricted model in all groups. Moreover, if one finds a misfitting item (or class) one has to free the restriction on the parameters of this item (or class) and to rerun the analysis to consider the new parameter estimates and the new model fit indices, and then to look for the next nonfitting item or class, etc. This is not feasible for a large number of countries and items. For a large number of countries a second strategy is more easily realizable.

Strategy II

In the second strategy, a multigroup LCA is conducted in which it is assumed that there is full measurement invariance across countries and the number of classes does not differ across countries. That means that a model with full measurement invariance is enforced. However, as we explain below the result of such an analysis could be a model with non-invariance or with partial invariance. The best fitting model is selected according to an exploratory strategy (e.g. applying information criteria). Using this strategy leads to more classes than one would usually find in a single group (country) but less classes than one would obtain in an analysis with only country-specific classes. This strategy also has the advantage that there is a higher power to detect small classes that exist in several countries but that would not be detected in country-specific analyses because their size within a country might be too small. Moreover, it could happen that some classes in some countries have a size of 0, which means that these classes do not exist in these countries. The basic idea of this strategy can be explained with respect to the case of two countries with three classes within each country. If the assumption of full measurement invariance holds, a solution with three measurement invariant classes will be expected. If there is no measurement invariance at all, a solution with six "measurement-invariant" classes will be

expected, whereas three classes have a size of 0 in the first country, and the remaining three classes have a size of 0 in the other country. That means that all three classes are not measurement invariant but specific to a country (leading to six classes using Strategy II). If two classes are measurement invariant, for example, a solution with four classes will be expected with one class having a size of 0 in one country and another class having a size of 0 in the other country. We illustrate this strategy with respect to the TALIS 2013 data.

Multilevel Latent Class Analysis

In multilevel latent class it is assumed that the groups (level-2 units) are randomly selected from a population. In the most general model, the parameters of a latent class model can differ between the level-2 units resulting in a model that is equivalent to a non-restricted multigroup latent class model (Vermunt, 2003[84]). However, with many level-2 units and small sizes of level-2 units the analysis of such a model would lead to many parameters to be estimated and unstable parameter estimates (Vermunt, 2003[84]). Therefore, in multilevel LCA restrictions are put on the parameters to avoid these estimation problems. In particular, it is assumed that the parameters stem from a certain distribution. When formally defining the model (see Annex 5.A), we describe this idea in more detail.

In order to determine the appropriate number of classes, information criteria can be computed. Lukočienė, Varriale and Vermunt (2010[99]) scrutinised the behaviour of several information criteria in the context of multilevel LCA. They came to the conclusion that when one wants to determine the number of latent classes at the higher level, the most appropriate sample size for the BIC and the CAIC is the higher level sample size (number of higher level units). They found that the BIC, CAIC and Information Complexity (ICOMP) behave well when the number of individuals per group is large ($n_j \geq 15$), whereas AIC3 performs better when the number of individuals per group is small ($n_j = 5$). Given the large sample sizes in the TALIS study we decided to consider the BIC as selection criterion for the number of classes in the application presented. For multilevel LCA bootstrap measures of goodness of fit are not available.

Model Evaluation and Fit Statistics for Measurement Invariance Testing

There are two ways to compare an unrestricted latent class model with a latent class model with full or partial measurement invariance: the likelihood ratio difference test and information criteria.

Likelihood ratio difference test

The likelihood ratio test can be applied to compare an unrestricted latent class model with a latent class model with full or partial measurement invariance because the latter is nested within the former. However, the likelihood ratio difference test requires large samples and is not valid in the situation of sparse tables. In the case of sparse tables, the bootstrap likelihood ratio test (BLRT) can be applied. The BLRT, however, is time consuming and is not applicable when the estimation of a model is very time consuming. In the case of many countries and no restrictions (many parameters to be estimated) the analysis of a model can take many hours depending on the computers available. In such a case a resampling procedure is not feasible, in particular, if one wants to compare several models. Moreover, bootstrap model fit criteria are not available for multilevel models.

Information criteria

The information criteria can be easily estimated for different models differing in the restrictions on the parameters (no restrictions, partial measurement invariance, full measurement invariance) and the number of classes. The best fitting model is the model with the lowest value of an information criterion. Whereas the AIC-3 seems to be appropriate for small sample sizes, the BIC seems to be appropriate for large sample sizes and multilevel models. Given the large sample sizes of large-scale assessment studies, the use of the BIC is recommended.

Empirical Example, Practical Advice and Recommendations

In the following we will present an empirical application and discuss some practical issues and recommendations.

Application of LCA to the TALIS Data Set: School Participation

We illustrate the application of multigroup and multilevel LCA using the five items of the school participation scale presented above. The data stem from the TALIS 2013 data set. The subdata set consists of ratings of n = 7 436 principals from J = 38 countries. The countries differ in the sample sizes, from a minimum of 98 principals to a maximum of 1 070 principals (with most samples including between 150 and 200 principals). We will illustrate Strategy II presented on p. 75 because there are many countries. The analyses were done in the following way.

We used the computer programme Latent GOLD Version 5.1 (Vermunt and Magidson, 2016[86]). We created the basic syntax using the syntax generator. A critical issue of LCA is the number of starting values. The syntax generator generates 16 sets of starting values. We reran the same analysis several times using the default value of 16 sets of starting values and found that we got slightly different solutions with respect to the maximum of the likelihood. We increased the number of sets of starting values to 40 and found that this worked fine. As a practical guideline, it is important to check whether the number of starting values is sufficient. In complex data structures such as in the OECD data sets, it is advisable to increase the number of starting values and not to use the default value. Different numbers of starting values can be checked and it can be analysed what the minimum value starting values is by trying different numbers. Moreover, we added "identification iteration details" to get an identification check. This should always be done, because a model with few items and many classes might not be identified.

We conducted a multigroup LCA with measurement invariance across countries (invariant conditional response probabilities across countries, but non-invariant class sizes). We increased the number of classes until the BIC coefficient indicated the best fitting model, which was a model with six latent classes. The BIC coefficients for the first seven classes are depicted in Table 5.1. The part of the Latent GOLD syntax defining the model is given in Table 5.2. This model allows interpreting the classes and considering their distribution across the different countries. One obtains for each country a frequency distribution of the latent classes. To avoid redundancy we will not present the results of this model, but the results of the extension of this model to a two-level model.

Table 5.1. BIC coefficients for a multigroup latent class analysis about stakeholder participation in decision-making with measurement invariance across countries

	Number of latent level-1 classes						
	1	2	3	4	5	6	7
BIC coefficient	63 912.61	58 902.68	57 464.65	56 535.28	56 506.61	**56 492.47**	56 652.33

Note: The lowest BIC value is in bold.

Table 5.2. Latent GOLD syntax for a multigroup latent class model with six classes being measurement invariant across countries

Syntax	Explanation
variables	
independent IDCNTRY nominal;	*country code, country is a nominal scale*
dependent TC2G22A nominal, TC2G22B nominal, TC2G22C nominal, TC2G22D nominal, TC2G22E nominal;	*five participation items defined as nominally scaled items*
latent	
Cluster nominal 6;	
equations	
Cluster <- 1 + IDCNTRY;	*class sizes depend on the country*
TC2G22A <- 1 + Cluster;	*response probabilities depend on the classes but not on the country*
TC2G22B <- 1 + Cluster ;	
TC2G22C <- 1 + Cluster ;	
TC2G22D <- 1 + Cluster ;	
TC2G22E <- 1 + Cluster ;	

In a next step, we analysed whether taking the items as ordinal variables and applying the restrictions of an ordinal measurement model (see Annex 5A) would result in a lower BIC value. Because this was not the case (BIC = 56 510.17), we considered the restrictions of the ordinal model as too restrictive, and treated the items as nominally scaled for the rest of the analyses.

Based on the results of the multigroup analysis with six latent classes we conducted a two-level LCA (principals nested within countries) and analysed models with six classes on level 1 and an increasing number of level-2 classes. We increased the number of starting values to 600. Now we considered the BIC based on the number of groups as selection criterion. The BIC values are presented in Table 5.3. The fit improved until six classes. Because the results became unstable for more than six latent classes and were affected by identification problems, we took the model with six classes on level 2 as the best fitting model. The syntax for this model is shown in Table 5.4.

Table 5.3. BIC coefficients for a multilevel latent class analysis about stakeholder participation in decision-making with measurement invariance across countries

	Number of latent level-2 classes					
	1	2	3	4	5	6
BIC coefficient	57 317.82	56 023.21	55 596.97	55 381.76	55 240.92	**55 151.94**

Note: The lowest BIC value is in bold.

Table 5.4. Latent GOLD syntax for a multigroup latent class model with six classes on both levels

Syntax	Explanation
variables	
independent IDCNTRY nominal;	*country code, country is a nominal scale*
dependent TC2G22A nominal, TC2G22B nominal, TC2G22C nominal, TC2G22D nominal, TC2G22E nominal;	*five participation items defined as nominally scaled items*
latent	
GClass group nominal 6,	*six latent classes on level 2 are considered*
Cluster nominal 6;	*six latent classes on level 1are considered*
equations	
GClass <- 1;	*sizes of the classes are estimated*
Cluster <- 1 + GClass;	*classes on level 1 depend on the level-2 classes*
TC2G22A <- 1 + Cluster;	*response probabilities depend on the level-1 classes but not on the level-2 classes*
TC2G22B <- 1 + Cluster ;	
TC2G22C <- 1 + Cluster ;	
TC2G22D <- 1 + Cluster ;	
TC2G22E <- 1 + Cluster ;	

The conditional response probabilities for the level-1 classes are presented in Table 5.5.

The conditional probabilities for belonging to a level-1 class given a level-2 class are presented in Table 5.6.

In order to consider the distribution of the level-2 classes across countries, countries were assigned to level-2 classes using the assignment probabilities for the different classes. The modal assignment probability is 1, showing that the countries are very distinct and that the latent-2 classes represent national differences very well. Table 5.7 presents the cross-classifications of the 38 countries and the level-2 classes to which they were assigned.

Table 5.5. Latent GOLD output presenting the class-specific conditional response probabilities for the level-1 classes in the two-level model

	Level-1 Classes (Clusters)						
	1	2	3	4	5	6	Overall
Size	0.3066	0.1603	0.0895	0.2385	0.1002	0.1048	
TC2G22A	**Staff**						
Strongly disagree	0.0006	0.0154	0.0010	0.0002	0.0017	0.0000	0.0030
Disagree	0.0009	0.0768	0.0026	0.0000	0.0497	0.0000	0.0178
Agree	**0.8663**	**0.8140**	0.0396	**0.6548**	**0.8248**	0.1203	**0.6511**
Strongly agree	0.1322	0.0938	**0.9568**	**0.3450**	0.1237	**0.8796**	**0.3282**
TC2G22B	**Parents or guardians**						
Strongly disagree	0.0000	0.0394	0.0000	0.0000	0.0209	0.0037	0.0088
Disagree	0.0109	**0.6444**	0.0000	0.0235	**0.3993**	0.0210	0.1545
Agree	**0.9730**	**0.3108**	0.1191	**0.9328**	**0.5723**	**0.4181**	**0.6825**
Strongly agree	0.0161	0.0054	**0.8809**	0.0437	0.0075	**0.5572**	0.1543
TC2G22C	**Students**						
Strongly disagree	0.0000	0.0442	0.0046	0.0007	0.0579	0.0000	0.0135
Disagree	0.0935	**0.5833**	0.0130	0.1136	**0.5764**	0.0201	0.2103
Agree	**0.8839**	**0.3654**	**0.3500**	**0.8581**	**0.3549**	**0.5402**	**0.6577**
Strongly agree	0.0227	0.0072	**0.6324**	0.0276	0.0108	**0.4397**	0.1185
TC2G22D	**Principal makes important decisions on her/his own**						
Strongly disagree	0.0003	0.1743	**0.6838**	**0.4811**	0.0077	0.0021	0.2050
Disagree	**0.5806**	**0.5799**	0.2529	**0.5130**	0.0877	**0.4226**	**0.4691**
Agree	**0.3698**	0.2246	0.0435	0.0059	**0.6624**	**0.4069**	0.2637
Strongly agree	0.0493	0.0211	0.0199	0.0000	0.2421	0.1684	0.0622
TC2G22E	**Collaborative school structure / mutual support**						
Strongly disagree	0.0016	0.0243	0.0112	0.0051	0.0000	0.0000	0.0066
Disagree	0.0301	0.1239	0.0260	0.0523	0.0421	0.0098	0.0491
Agree	**0.8034**	**0.6815**	**0.3766**	**0.6408**	**0.6829**	**0.3146**	**0.6435**
Strongly agree	0.1649	0.1703	**0.5863**	**0.3018**	0.2750	**0.6756**	**0.3007**

Note: Conditional probabilities larger than .30 are in bold.

Table 5.6. Latent GOLD output presenting the class-specific conditional probabilities for the level-1 classes (cluster) given the level-2 classes in the two-level model

	Level-2 Classes (GClasses)					
Level-1 Classes	1	2	3	4	5	6
1 – SAP+	**0.5879**	0.1935	0.0753	**0.4618**	0.2361	0.0115
2 – MStP-	0.0168	**0.4449**	0.0729	0.1200	0.2283	0.0012
3 – VSAP-	0.0627	0.0226	0.2697	0.0348	0.1097	0.0001
4 – SAP-	0.0616	0.1964	**0.5808**	0.0963	**0.3724**	0.0001
5 - MStWPaP+	0.0244	0.0868	0.0001	0.1376	0.0004	**0.8651**
6 – VSAP+	0.2465	0.0558	0.0012	0.1495	0.0530	0.1220

Note: Response probabilities larger than .30 are in bold.

Table 5.7. Classification of the 38 countries in the Level-2 classes

	Level-2 Classes (GClasses)					
	1	2	3	4	5	6
Australia	0	0	0	0	123	0
Brazil	0	0	1070	0	0	0
Bulgaria	0	0	0	197	0	0
Chile	0	0	0	178	0	0
Croatia	0	0	0	199	0	0
Cyprus*	0	0	0	98	0	0
Czech Republic	0	0	0	220	0	0
Denmark	0	0	0	148	0	0
Estonia	0	0	197	0	0	0
Finland	0	146	0	0	0	0
France	0	0	0	0	204	0
Georgia	194	0	0	0	0	0
Iceland	0	0	0	129	0	0
Israel	0	195	0	0	0	0
Italy	0	194	0	0	0	0
Japan	0	0	0	0	0	192
Korea	177	0	0	0	0	0
Latvia	116	0	0	0	0	0
Malaysia	0	0	0	0	0	150
Mexico	0	0	0	0	187	0
Netherlands	0	0	0	127	0	0
New Zealand	0	0	0	0	163	0
Norway	0	0	0	0	145	0
Poland	195	0	0	0	0	0
Portugal	0	0	185	0	0	0
Romania	0	0	197	0	0	0
Russia	198	0	0	0	0	0
Serbia	191	0	0	0	0	0
Singapore	0	159	0	0	0	0
Slovak Republic	0	0	0	193	0	0
Spain	0	0	192	0	0	0
Sweden	0	186	0	0	0	0
United States	0	0	0	0	122	0
Sub-national entities						
Abu Dhabi (United Arab Emirates)	0	0	0	0	166	0
Alberta (Canada)	0	0	0	0	182	0
England (United Kingdom)	0	0	0	0	154	0
Flemish Community (Belgium)	0	0	0	168	0	0
Shanghai-China	0	0	199	0	0	0
Sum	1071	880	2040	1657	1446	342

Notes: Numbers refer to the number of principals (schools), countries are assigned to classes based on the modal assignment probability. Response probabilities larger than .30 are in bold.

* *Note by Turkey:*

The information in this document with reference to 'Cyprus' relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey will maintain its position concerning the 'Cyprus issue'.

Note by all the European Union Member States of the OECD and the European Union:

The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

The level-1 classes (called "clusters" in the output) can be interpreted based on the conditional response probabilities. All schools are characterised by a collaborative school culture and mutual support (Item TC2G22E), but the classes differ with respect to the participation of the different groups and can be characterised as in Table 5.8.

Table 5.8. School-level classes for types of distributed leadership (Level-1 classes)

Based on principal reports

	Size (%)	Description
Class 1: SAP+	30.66	Strong participation of all groups, but it is not unlikely that the principal makes decisions alone.
Class 2: MStP-	16.03	Mainly staff members participate in decisions, and it is rather unlikely that the principal makes decisions alone.
Class 3: VSAP-	8.95	Very strong participation of all different groups, rather unlikely that a principal makes decisions alone.
Class 4: SAP-	23.85	Strong participation of all groups, unlikely that the principal makes decisions alone.
Class 5: MStWPaP+	10.02	Mainly staff members participate in decisions, parents to a certain degree and the probability that students can participate is rather low, it is likely that principal makes decisions alone
Class 6: VSAP+	10.48	Very strong participation of all groups, but it is rather likely that the principal makes important decisions alone.

Note: Please refer to the text for an explanation of class names (abbreviations).

The classes can be ranked according to the participation structure in the following way

- Very strong participation: Classes 3 and 6. The classes 3 and 6 are classes with very strong participation of all groups, but they differ in the role of the principal, in class 6 it is likely that she/he will make also important decisions alone, but this is unlikely in class 3. We will use the abbreviations VSAP- (very strong participation of all groups, P-: principal does not make decisions alone) for Class 3 and VSAP+ for Class 6. Together, 19.43% of all schools belong to very strong participation schools.
- Strong participation: Classes 1 and 4. The classes 1 and 4 are classes with strong participation of all groups, but they differ in the role of the principal, in class 1 it is not unlikely that she/he will make also important decisions alone, but this is very unlikely in class 3. We will use the abbreviations SAP+ (strong participation of all groups, P-: principal does not make decisions alone) for class 1 and SAP1- for class 4. Together, 54.51% of all schools belong to strong participation groups.
- Strong participation of staff, weak participation of other groups, mainly parents: Class 5. We abbreviate this class as MStWPaP+ (mainly staff, weaker parents, P+: principal makes decisions alone). 10.02% of all schools belong to this class.
- Strong participation of staff, but much weaker participation of parents (and students): Class 2. We abbreviate this class as MStP- (mainly staff, P-: principal does not make decisions alone). In contrast to Class 5, the probability that parents have the opportunity to participate is smaller, and the principal does not make decision on her or his own. 16.03% of all schools belong to this class of lowest participation opportunities.

Based on the meaning of the level-1 classes, the level-2 classes can be interpreted. In Table 5.6 the conditional probabilities for belong to a level-1 class (cluster) for members of a level-2 class (GClass) are presented. Table 5.9 shows the meaning and countries assigned to level-2 classes (also see Table 5.7).

Table 5.9. Country-level classes for types of distributed leadership (Level-2 classes)

	Description	Countries
Class 1	Members of this class have a high probability to belong a level-1 class characterised by strong participation of all and a comparatively high probability for a principal making important decisions on her or his own.	Georgia, Korea, Latvia, Poland, Russia, Serbia
Class 2	About 44.49% of schools in this class belong to a level-1 class characterised by low participation opportunities for parents/guardians, but strong participation opportunities for staff, and principals not deciding alone. About 39% of all schools a characterised by strong participation opportunities for all, half of them with principals showing a comparatively high probability of making decisions alone, half of them not.	Finland, Israel, Italy, Singapore, Sweden
Class 3	Members of this class belong predominantly to latent-1 classes with (very) strong participation opportunities for all, and principals not deciding alone.	Brazil, Estonia, Portugal, Romania, Shanghai (China), Spain
Class 4	Mainly schools with strong participation opportunities for all and principals showing a comparatively high probability of making decisions on their own (46.18%). Other schools almost equally distributed over other types with the exception of schools with very strong opportunities for all and principals not deciding alone.	Bulgaria, Chile, Croatia, Cyprus[2], Czech Republic, Denmark, Flemish Community (Belgium), Iceland, Netherlands, Slovak Republic.
Class 5	Heterogeneous class with many types. 37.24% of schools belong to strong participation of all and principals not deciding alone. Low probabilities for MStWPaP+ and VSAP+.	Abu Dhabi, Alberta (Canada), Australia, England (United Kingdom), France, Mexico, New Zealand, Norway, United States
Class 6	Members of this class belong primarily to level-1 class 5, characterised by strong participation possibilities for staff, weaker possibilities for parents (and low for students) and principals making decisions on their own.	Japan, Malaysia

Practical Advice and Recommendations

The application of multigroup and multilevel LCA was illustrated with an example from the TALIS study comprising different principals (schools) in different countries. In the application presented, multigroup LCA allows to consider the distribution of the different classes in the different countries in detail, which is a more fine-grained rendering of the data. Multilevel analyses group countries into different clusters showing the same distribution of level-1 classes. This is less fine-grained but facilitates the interpretation of the data a lot, as only six level-2 classes (instead of 38 country-specific distributions) have to be considered.

Two-level LCA can also be applied to teacher and student data. In this case, teacher/students (level 1) would be nested in schools (level 2). The different countries can be represented by dummy variables and included as predictor variables for the latent class variables. Based on this example we will discuss some general issues and give some practical advice.

[2] *Note by Turkey*:

The information in this document with reference to 'Cyprus' relates to the southern part of the Island. There is no single authority representing both Turkish and Greek Cypriot people on the Island. Turkey recognises the Turkish Republic of Northern Cyprus (TRNC). Until a lasting and equitable solution is found within the context of the United Nations, Turkey will maintain its position concerning the 'Cyprus issue'.

Note by all the European Union Member States of the OECD and the European Union:

The Republic of Cyprus is recognised by all members of the United Nations with the exception of Turkey. The information in this document relates to the area under the effective control of the Government of the Republic of Cyprus.

Strategy I versus Strategy II

We illustrated Strategy II because of the many countries and because this strategy has to our knowledge not been illustrated in detail in the literature before. Using Strategy I has some limitations in the current application. First, analyses in 38 countries would have been done separately to figure out how many latent classes are required in the different countries. If the assumption of full measurement invariance has to be rejected for all 38 countries – which is usually the case – all 38 countries will have to be compared to each other with respect to the latent classes and items to check whether a model with partial measurement invariance would fit the data. That means that one has to do 38*37/2 = 703 comparisons (country by country) to figure out whether there are countries being more similar and less similar in the class-specific response patterns. This is very cumbersome if not impossible. Using Strategy II countries can be clustered with respect to their class structures in a data-analytic way. We found six clusters of countries, a small number that simplifies the interpretation of similarities and differences between the countries a lot.

If there are only very few countries that should be compared, we strongly recommend Strategy I that is well explained and illustrated in Eid and Diener (2001[98]), Eid, Langeheine and Diener (2003[81]) and Kankaraš, Moors and Vermunt (2018[82]).

How to Deal with Violations of Measurement Invariance

In multinational studies measurement invariance seldom holds. How to deal with violations of measurement invariance? LCA offers several ways to model partial measurement invariance that have been presented. For example, it can be analysed whether there are groups of countries that are measurement invariant and groups of countries that differ. If the measurement in-equivalence is due to single items or countries one can think about eliminating these items or countries. However, this should only be done if there are strong reasons, for example, if the translation of an item was ambiguous or participants in a country did not follow the instructions. This cannot be decided on a statistical analysis alone and needs further information. In general, items and countries do not have to be excluded, because LCA is flexible enough to consider aberrant items and countries by relaxing the assumption of full measurement invariance.

Critical Issues

In order to apply LCA, the items should be correlated because the latent class structure explains the correlations of the observed items. If the different items were not correlated, it would not make sense to apply LCA to the data set. LCA has its limits when there are many items having many response categories and being very weakly correlated. In this case, many classes would have to be considered, and the comparison of many countries would be bothersome.

An important critical issue is the selection of the number of classes. In the application presented we decided for six latent classes and a model for nominal observed variables. However, the BIC value for five classes was only slightly higher, also the model with ordinal restrictions had only a slightly higher BIC value. Hence, a model with five classes might also fit the data well. In order to decide about the number of classes, several class solutions can be considered and the choice of the model can also depend on theoretical considerations. For example, in the current application the solutions with five and six classes can be compared. If the sixth class does not strongly differ from other classes, is not informative at all or very small, one can prefer a five-class solution. The smallest class

in our solution (Class 3 in Table 5.5 and Table 5.8) shows a very interesting response pattern differing from the other response pattern. Therefore, a six class solution was chosen. One might also prefer a model with ordering restrictions on the categories for theoretical reasons. Then, one would opt for the model with order restrictions. However, interpreting the response probabilities for the different categories in Table 5.5 is the more basic approach and does not require ordering restrictions.

Software

Multigroup analyses can be done with the computer programmes Latent Gold (Vermunt and Magidson, 2016[86]), LCCA (Schafer and Kang, 2013[100]), LEM (Vermunt, 1997[101]), M*plus* (Muthén and Muthén, 1998-2017[25]) and mdltm (von Davier, 2005[87]; von Davier, 2010[88]; von Davier and Rost, 2016[89]). Multilevel LCA models can be analysed with Latent Gold and M*plus*. The programmes LCCA and LEM are non-commercial programmes (open access). However, both programmes do not have bootstrap facilities. Latent Gold and M*plus* are the more general programmes. M*plus* only provides a bootstrap likelihood ratio difference test for comparing models differing in the number of classes. Latent Gold offers more bootstrap model fit facilities, for example, a bootstrap likelihood ratio test for testing a LCA model confirmatorily, and a bootstrap likelihood ratio difference test for comparing an unrestricted model with a model with full or partial measurement invariance.

Comparative Overview

LCA is a model for categorical observed variables measuring categorical latent variables. In this regard, LCA is unique and differs from all other approaches presented in the other chapters. The choice of a model should be based on theoretical considerations. Whenever it is assumed that a construct is typological and not dimensional in nature, LCA is the approach of choice. LCA requires that the observed variables are categorical. In the case of metrical observed variables other approaches have to be applied. LCA can also be applied when assumptions of other models for categorical observed variables such as latent trait models are violated. For example, in contrast to latent trait models LCA does not require that items are ordered on one or more latent dimensions (although a model with ordered latent classes can also be applied) (Heinen, 1996[72]; Kankaraš, Moors and Vermunt, 2018[82]). In contrast to other approaches, LCA does not assume that a country is homogeneous with respect to the parameters of a model, but that there can be subgroups within countries that are not equivalent with respect to the parameters of a model. These subgroups can be compared across countries. Hence, LCA is a very flexible approach for international studies.

Annex 5.A. Formal Definition of the Models

In order to define the LCA in a formal way we follow the notation of Vermunt (2003[84]).

Latent Class Model

The classical latent class model is defined by the following equation (Vermunt, 2003, p. 216[84]):

$$P(\boldsymbol{Y}_i = \boldsymbol{s}) = \sum_{t=1}^{T} P(X_i = t) \cdot P(\boldsymbol{Y}_i = \boldsymbol{s} | X_i = t)$$

$$= \sum_{t=1}^{T} P(X_i = t) \prod_{k=1}^{K} P(Y_{ik} = s_k | X_i = t)$$

Equation 5.1

with

- Y_{ik}: observed response variable
  - i: individual, $i = 1, \ldots, n$
  - k: item; $k = 1, \ldots, K$
  - s_k: category of item k; $s_k = 1, \ldots, S_k$
- $\boldsymbol{Y}_i$: full vector of responses of individual i
- $\boldsymbol{s}$: possible response pattern
- X_i: latent class variable
  - t: latent class, $t = 1, \ldots, T$
- $P(\boldsymbol{Y}_i = \boldsymbol{s})$: Probability of an observed response pattern $\boldsymbol{s}$
- $P(X_i = t)$: Probability of a latent class t (class size)
- $P(\boldsymbol{Y}_i = \boldsymbol{s} | X_i = t)$: Class-specific (conditional) probability of an observed response pattern $\boldsymbol{s}$
- $P(Y_{ik} = s_k | X_i = t)$: Class-specific (conditional) probability of an observed response s_k

The equation $P(\boldsymbol{Y}_i = \boldsymbol{s} | X_i = t) = \prod_{k=1}^{K} P(Y_{ik} = s_k | X_i = t)$ defines the conditional independence of the observed responses (local independence).

The class probabilities and the conditional response probabilities can also be written in form of the logit parameterisation that is used for extending the standard model to a multigroup and multilevel model:

$$P(X_i = t) = \frac{\exp(\gamma_t)}{\sum_{r=1}^{T} \exp(\gamma_r)}$$

Equation 5.2

$$P(Y_{ik} = s_k|X_i = t) = \frac{\exp(\beta_{s_k t}^k)}{\sum_{r=1}^{S_k} \exp(\beta_{rt}^k)}$$ **Equation 5.3**

For identification reasons a constraint has to be imposed on each equation, e.g., $\gamma_1 = 0$ and $\beta_{1t}^k = 0$.

Definition of the latent class model for ordinal variables

So far, we have assumed that the observed variables are measured on a nominal scale. Often, however, the categories of a response scale are ordered like the rating scale in the TALIS example presented in Table 5.5. The LCA model can be extended to ordinal response scales by putting restrictions on the response categories. The advantage is that this saves parameters to be estimated and the estimated class-specific (conditional) mean values can be interpreted. This could simplify the presentation of the results because the conditional probability distribution of the response categories might not have to be presented. In the case of ordinal response variables the parameters $\beta_{s_k t}^k$ in Equation 5.3 are restricted in the following way (Vermunt and Magidson, 2016, p. 23[86]):

$$\beta_{s_k t}^k = \beta_{s_k 0}^k + \beta_{0t}^k \cdot y_{s_k}^{k*}$$ **Equation 5.4**

where $y_{s_k}^{k*}$ are the category scores, resulting in an adjacent-category ordinal logit measurement model. The meaning of the parameters depends on how the category scores $y_{s_k}^{k*}$ are chosen. If the first category, for example, is assigned a score of 0, one gets a so-called baseline-category logit. Another possibility would be to choose an effect coding of the categories of the observed variables (Vermunt and Magidson, 2016, p. 16[86]).

Multigroup Latent Class Analysis

A multigroup latent class model is defined by adding an index *j* for the group (Vermunt, 2003, p. 216[84]):

$$P(\boldsymbol{Y}_{ij} = \boldsymbol{s}) = \sum_{t=1}^{T} P(X_{ij} = t) \cdot P(\boldsymbol{Y}_{ij} = \boldsymbol{s}|X_{ij} = t)$$
$$= \sum_{t=1}^{T} P(X_{ij} = t) \prod_{k=1}^{K} P(Y_{ijk} = s_k|X_{ij} = t)$$ **Equation 5.5**

with

- Y_{ijk}: observed response variable
  - i: individual within group j; $i = 1, \ldots, n_j$
  - j: group; $j = 1, \ldots, J$
  - k: item; $k = 1, \ldots, K$

- s_k: category of item k; $s_k = 1, \ldots, S_k$

- $\boldsymbol{Y}_{ij}$: full vector of responses of individual i in group j
- $\boldsymbol{s}$: possible response pattern
- X_{ij}: latent class variable
  - T: latent class, $t = 1, \ldots, T$
- $P(\boldsymbol{Y}_{ij} = \boldsymbol{s})$: Probability of an observed response pattern $\boldsymbol{s}$
- $P(X_{ij} = t)$: Probability of a latent class t (class size)
- $P(\boldsymbol{Y}_{ij} = \boldsymbol{s} | X_{ij} = t)$: Class-specific (conditional) probability of an observed response pattern $\boldsymbol{s}$
- $P(Y_{ijk} = s_k | X_{ij} = t)$: Class-specific (conditional) probability of an observed response s_k

The equation $P(\boldsymbol{Y}_{ij} = \boldsymbol{s} | X_{ij} = t) = \prod_{k=1}^{K} P(Y_{ijk} = s_k | X_{ij} = t)$ defines the conditional independence of the observed response (local independence).

In form of the logit parameterisation, the class probabilities and the conditional response probabilities are defined in the following way:

$$PP(X_{ij} = t) = \frac{\exp(\gamma_{tj})}{\sum_{r=1}^{T} \exp(\gamma_{rj})} \qquad \textbf{Equation 5.6}$$

$$PP(Y_{ijk} = s_k | X_{ij} = t) = \frac{\exp(\beta_{s_k tj}^{k})}{\sum_{r=1}^{S_k} \exp(\beta_{rtj}^{k})} \qquad \textbf{Equation 5.7}$$

According to Equation 5.6 the size of the latent classes can differ between groups, and according to Equation 5.7 the conditional response probabilities can be group-specific. Full measurement invariance means that the conditional response probabilities $P(Y_{ijk} = s_k | X_{ij} = t)$ do not differ between groups. Consequently, also the parameters $\beta_{s_k tj}^{k}$ do not differ between groups and have to be set equal across groups.

Kankaraš et al. (2018[82]) decompose a parameter $\beta_{s_k tj}^{k}$ according to a logistic regression into two parameters that we will notate with $\alpha_{s_k j}^{k}$ and $\alpha_{s_k tj}^{k}$: $\beta_{s_k tj}^{k} = \alpha_{s_k j}^{k} + \alpha_{s_k tj}^{k}$. The first parameter ("intercept") does not depend on the class (but on the group) and characterises the difficulty of an item. The second parameter ("slope") depends on the class and the group and represents the strength of relationship between the latent and the observed variable. In the case of full measurement both parameters do not depend on the group. If the assumption of full measurement invariance has to be rejected, the decomposition into the two parameters allows testing specific hypotheses about the reason of the violation. If, for example, only the parameters $\alpha_{s_k j}^{k}$ differ between groups but not the parameters $\alpha_{s_k tj}^{k}$, then the groups differ only in the difficulties ("intercepts") but not the associations with the classes ("slopes"). According to Kankaraš et al. (2018[82]) such a model is similar to the concept of metric equivalence in models for continuous variables such as multigroup confirmatory factor analysis. Kankaraš et al. (2018[82]) discuss and illustrate with empirical

examples how different hypotheses about group differences can be tested based on this decomposition.

Definition of the multigroup latent class model for ordinal variables

In a multigroup LCA for ordinal response variables one obtains an equation analogous to Equation 5.4 by adding the index *j* for the group:

$$\beta^k_{s_k tj} = \beta^k_{s_k 0j} + \beta^k_{0tj} \cdot y^{k*}_{s_k}$$ **Equation 5.8**

Multilevel Latent Class Analysis

The general formula of the model (both for nominal and ordinal observed variables) is the same as for the multigroup case. In Equation 5.5 individuals *i* are now considered level-1 units and groups *j* are considered level-2 units. Vermunt (2003[84]) distinguishes between two different approaches of multilevel LCA, a parametric and a nonparametric approach.

Parametric multilevel latent class analysis

In the parametric multilevel LCA model it is assumed that the group-specific effects (effects of level-2 units) stem from a parametric distribution. If one assumes, for example, that the conditional response probabilities are invariant across level-2 units (measurement invariance) but the class sizes differ between level-2 units, the group-specific parameters in Equation 5.6 can be restricted in the following way (Vermunt, 2003, p. 218[84]):

$$\gamma_{tj} = \gamma_t + \tau_t \cdot u_j$$ **Equation 5.9**

with the assumption that the level-2-specific effects stem from a standard normal distribution [$u_j \sim N(0,1)$], and with one identifying constraint on each of the two parameters γ_t and τ_t such as $\gamma_1 = 0$ and $\tau_1 = 0$. If one makes this identifying constraint, a parameter contrasts the size of class *t* with the first class ($t = 1$) that is taken as reference class. In Equation 5.9it is assumed that the random effects for the different latent classes are unidimensional, an assumption that might be too strong and could be relaxed to a multidimensional structure (Vermunt, 2003[84]). Based on Equation 5.9 for each non-reference class an intraclass-correlation can be computed:

$$r_{It} = \frac{\tau_t^2}{\tau_t^2 + \pi^2/3}$$ **Equation 5.10**

with $\pi = 3.14$.

Nonparametric multilevel latent class analysis

The parametric multilevel LCA assumes that the random level-2 effects stem from a normal distribution which is rather restrictive. Vermunt (2003[84]) has developed a nonparametric multilevel LCA approach that defines latent classes on level 2. This approach allows clustering level-2 units, for example, countries according to the sizes of the level-1 classes (that are invariant across countries). The level-2 latent class variable is denoted by *W*, and

W_j is a value of group j on the latent class variable. If m stands for a specific latent class ($m = 1, \ldots, M$) with size π_m, the probability that a member of level-2 class m belongs to level-1 class t is:

$$P(X_{ij} = t | W_j = m) = \frac{\exp(\gamma_{tm})}{\sum_{r=1}^{T} \exp(\gamma_{rm})} \quad \textbf{Equation 5.11}$$

It is possible to decompose γ_{tm} in $\gamma_{tm} = \gamma_t + u_{tm}$, but u_{tm} does not come from a specified distribution as the random effects in the parametric approach do. In illustrating the use of multilevel LCA for OECD data we refer to the nonparametric approach because we do not want to make the assumption that the countries can be ordered on a latent dimension.

If one applies multilevel LCA to the situation that, for example, teachers (level-1 units) are nested within schools (level-2 units) belonging to different countries, countries can be considered as a level-2 covariate. The different countries have to be represented by dummy variables. For one level-2 predictor Z_{1j} (e.g. one dummy variable), Equation 5.11 can be extended in the following way:

$$P(X_{ij} = t | Z_{1j}, W_j = m) = \frac{\exp(\gamma_{0tm} + \gamma_{1t} Z_{1j})}{\sum_{r=1}^{T} \exp(\gamma_{0rm} * + \gamma_{1r} Z_{1j})} \quad \textbf{Equation 5.12}$$

The extension to more than one level-2 predictor is straightforward. It is also possible to include level-1predictor variables (Vermunt, 2003, p. 220[84]).

Chapter 6. Conclusion: An OECD conference on the Cross-cultural Comparability of Questionnaire Measures in Large-scale Assessments

Francesco Avvisati, Noémie Le Donné, Marco Paccagnella

The Organisation for Economic Co-operation and Development (OECD) is well known among education researchers for its role in promoting cross-country comparisons based on large-scale survey projects. The three largest OECD education surveys include the Programme for International Student Assessment (PISA), the Survey of Adult Skills (Programme for the International Assessment of Adult Competencies, or PIAAC), and a survey of teachers, also known as Teaching and Learning International Survey (TALIS); several other survey projects are currently in development. This report summarises the discussions from a methodological conference held in Paris on 8-9 November 2018, and which aimed at developing a common understanding and approach to the challenge of comparability of questionnaire responses and scales. We believe these issues to be relevant to survey programmes in other fields too, and to many researchers and practitioners using international survey data.[3]

Overview

The value of cross-country comparisons is at the heart of large-scale international surveys. But as household surveys expand from tools to measure objective attributes (age, household size) and behaviours (e.g. unemployment or job-seeking behaviours) to instruments to assess subjective attitudes (e.g. attitudes towards migrants, or subjective well-being), and as skills assessments aim no longer to measure only knowledge of mathematics, but also psychological traits such as perseverance, new challenges for the validity and comparability of survey results emerge, and old issues acquire renewed salience. Reflective latent constructs measured through self-reports, for example, are particularly affected by subtle linguistic differences in the translated questionnaires and by broader cultural differences. These may introduce variation in participants' understanding of survey questions, and therefore in the relationship between their responses and the target latent construct. Similarly, when confronted with Likert items, with generic frequency scales ("often", "sometimes", "never or almost never", …), or with subjective rating scales ("on a scale from 1 to 10"), cultural norms may mediate the response process of participants. As a result, international surveys may fall short of their objective to facilitate comparisons across countries.

Questions around cross-cultural comparability were recently the focus of a methodological conference at the OECD headquarters: How can different levels of comparability be defined? How can they be identified in the data? How should violations of comparability be addressed when analysing and reporting these data, to prevent misuse of the data in policy discussions? How can instruments be designed in order to maximise comparability?

The conference brought together leading experts in questionnaire design and in the statistical modelling of survey responses with representatives from the industry involved in the development of questionnaires, data products and reports. It aimed at identifying

[3] The authors take full responsibility for the information provided. This summary is not a consensus document approved by all conference participants

areas where current practices for designing and analysing questionnaires in cross-national large-scale surveys can improve, while keeping in mind the practical constraints, the timelines, and the reporting goals of such surveys.

The problem

Much effort in large-scale cross-national surveys is devoted to ensuring that the choice of particular item types, the questionnaire translations or their administration procedures do not introduce unintended bias in comparisons. Yet even the most rigorous application of preventive measures cannot guarantee the full comparability of measurement instruments (Davidov et al., 2014[9]). As an illustration, Lommen, van de Schoot and Engelhard (2014[102]) show how a particular questionnaire measure for post-traumatic stress symptoms in soldiers cannot be compared before and after their deployment in a war zone, despite the use of a within-subject design and the repeated administration of the same instruments under the same procedures.

Measurement invariance can be defined as a conditional independence property of the measurement model with respect to a set of sub-populations within the parent population (e.g. language groups, or gender, or time) (Mellenbergh, 1989[103]; Horn and Mcardle, 1992[104]; Meredith, 1993[8]).

With multiple indicators and known sub-populations, three classes of measurement models are often used. Multigroup confirmatory factor analysis (MGCFA) is the most popular approach when both the (latent) variable of interest and the manifest indicators (e.g. questionnaire responses) are continuous (or are treated as such, e.g. in the case of Likert scales). When the manifest indicators are ordinal (or categorical), Categorical MGCFA or item-response-theory (IRT) models can be used. When the latent variable is categorical, latent class analysis (LCA) models are appropriate.

Once combined with a particular measurement model, the assumption of measurement invariance can be formalised as a set of restrictions on model parameters. This allows to assess whether the assumption of measurement invariance holds, by either testing these restrictions in a frequentist hypothesis testing framework, or by comparing goodness of fit across models with or without these restrictions (e.g. in a Bayesian framework). For example, in a multigroup item-response theory (IRT) framework, the conditional independence assumption implies the lack of differential item functioning (DIF). In a MGCFA framework, conditional independence implies that a model with common factor loadings and intercepts for all groups fits the data as well as a model with group-specific parameters, once the estimation properly accounts for the random component in the data-generating process.

A standard of the past

The procedures for assessing measurement invariance within the framework of MGCFA are probably the best known and the closest to a current standard. Typically, three (nested) models are estimated. A “configural” model imposes the same configuration of zero and non-zero loadings for all groups, but allows all model parameters to vary across groups. A “metric” invariant model restricts item loadings to be common across groups, but allows item intercepts to vary freely. A “scalar” invariant model, in line with the above definition of measurement invariance, restricts all model parameters (i.e. loadings and intercepts) to be common across groups (van de Schoot, Lugtig and Hox, 2012[105]; Davidov et al.,

2014[9]).[4] Model-fit indices are then compared across these nested models, and conclusions are drawn about whether the data conform to the stronger "scalar invariance" hypothesis, or to the weaker "metric" or "configural" invariance hypotheses.

There are multiple problems with the application of this procedure to large-scale international studies, as was repeatedly stated in meeting presentations. When the number of observations per group is small, likelihood ratio tests have limited power; while with large groups, violations of invariance detected in such tests may be inconsequential for the substantive inferences. More generally, the statistical tests involved have been developed in the case of two groups, i.e. when testing only one (set of) restriction(s) at a time: in this case, substantiated cut-off criteria exist. With a large number of groups, multiple hypotheses are tested simultaneously, and blind application of standard cut-off values can lead to systematic rejection of invariance, due to chance capitalisation. The problem is compounded by the fact that in realistic settings (when violations of measurement invariance may be due to cultural or language specificities), the hypotheses are not independent, neither across items, nor across groups. This has led to somewhat ad-hoc fixes such as using, instead of likelihood ratio tests, global model-fit measures whose sampling distributions are unknown, and determining the test cut-off values based on simulation studies. The use of these cut-offs in situations that differ, in meaningful ways (number of factors, groups, observations, etc.), from the simulation conditions under which they were derived is, however, not warranted (Rutkowski and Svetina, 2013[106]; Rutkowski and Svetina, 2016[107]). Moreover, the binary nature of the test still leaves practitioners with no idea about the extent to which misspecifications in the measurement model affect the secondary analyses of the latent trait, and the global nature of the test provides little information about the specific restrictions (groups and item parameters) that are responsible for the rejection.

In this situation, survey organisations may be tempted to increase the chances of instruments passing the tests by limiting participation to groups that are more similar or by including redundant items and limiting the variation in question types. The former strategy may severely limit the number of meaningful comparisons for many participants, as in reality, countries and cultures do not fall into clearly distinct groups; the latter strategy would result in sacrificing the validity gains that result from triangulating multiple perspectives and measures.

Perhaps more concerning is the fact that the most frequent practice is, in fact, to simply ignore the possible non-equivalence of measurement in cross-cultural research: many secondary users of the data compare respondents' answers and scale values derived from statistical models without acknowledging, and discussing, the potential threats to comparability (Boer, Hanke and He, 2018[108]). Other scholars resort to generalisations based on the analysis of single items – a situation in which comparability of measurements cannot be formally assessed based on the properties of a measurement model. Finally, when measurement invariance across countries has been rejected, many scholars move on to within-country analyses, without further assessing the measurement-invariance hypothesis with respect to subnational groups (in part, due to sample size limitations).

[4] A "strict" level of invariance can be defined when residual variance parameters are also restricted to be equal among groups.

Excitement around new developments

In recent years, many alternative paradigms in measurement-equivalence research have emerged. To what extent will these lead to the establishments of new standards in international large-scale surveys and support robust conclusions about cross-country differences?

Dealing with imperfect comparability of measurements when scaling and reporting continuous traits

The first sessions of the conference dealt with statistical approaches to analyse and report on data potentially affected by non-equivalence issues, in situations where the latent trait of interest is modelled as a continuous trait. The presenters and discussants in these sessions debated the merits of different models with application and simulation studies. This report does not provide a comprehensive textbook introduction to each of the statistical methods (though it includes some references for interested readers), but focuses, instead, on the contingencies and practicalities that emerged from these discussions.

Partial invariance

Model-building approaches are very common in the IRT framework, and have often been used by MGCFA practitioners in response to the failure to establish full scalar invariance. Starting from a fully invariant (scalar invariant) model, these approaches estimate item-level fit indices for every group; identify the items for which certain groups exhibit high level of misfit (usually referred to as differential item functioning, or DIF, in IRT), then deal with misfit by sequentially releasing constraints, until adequate fit is reached. This results in so-called "partial invariance" models, whereby the conditional independence holds for some measurements (often referred to as "anchor items"), but not all (Byrne, Shavelson and Muthén, 1989[6]; Steenkamp and Baumgartner, 1998[10]). This approach is currently in use in the PISA assessment, both in the scaling of the cognitive component (von Davier et al., 2018[109]) and in the analysis of questionnaire scales (Buchholz and Hartig, 2017[110]).

While several tools to detect problematic items are commonly used, participants were reminded of some caveats: statistical tests have limited power with small sample sizes (number of observations per item and group) and short scales; and item-level fit statistics are contingent on other items and on the distribution of the latent trait among respondents. The latter means, on the one hand, a certain path-dependency (dependence on prior decisions) in situations where multiple items are affected by misfit, and, on the other hand, that outlier-detection procedures may not work well for items whose locations do not overlap with the latent trait distribution.

Participants were also reminded that there is little guidance in the existing research literature regarding the more substantive question of whether meaningful comparisons of latent means can be conducted, in situations where only partial invariance holds. How many non-invariant items are required to build a "comparable" scale? What other criteria should be taken into account?

In this respect, a simulation study presented by Artur Pokropek contained some comforting results, showing that when the non-invariant items are correctly identified, a MGCFA model with just one invariant item out of five across 75% of the groups did recover latent group means reasonably well (Pokropek, Davidov and Schmidt, 2019[111]).

Alignment optimisation

In recent years, an alternative response to the failure to establish full scalar invariance in MGCFA has gained popularity, the so-called alignment optimisation approach (see Chapter 2). This approach tolerates small differences, even if there are many of them. The popularity of the approach is due to its simplicity and to its availability in the popular software package M*plus* (Muthén and Muthén, 1998-2017[25]). It requires only two steps: 1) estimation of a model with group-specific parameters ("configural model"); 2) minimisation of a loss function which depends on differences between parameters across groups, leading to a "rotated" solution which forces the group means and variances from the configural model on a same scale. The procedure is similar to factor rotation in exploratory factor analysis (EFA) (Asparouhov and Muthén, 2014[19]) and can equally be applied in the IRT context (Muthén and Asparouhov, 2014[20]). Matthias von Davier, in particular, also highlighted how the alignment optimisation method is very similar to the simultaneous test-linking approach proposed by Haberman (2009[112]).

While alignment optimisation has an intuitive practical appeal, including a simple explanation ("minimise differences between measurement-model parameters") and limited computational demand, participants at the conference were reminded of several drawbacks of the alignment method. To start, the method promises to make group means from configural models "most comparable", but there are no clear established criteria to determine if this solution is "comparable enough" to lead to meaningful comparisons of group means. In the simulation study presented by Artur Pokropek, latent means were recovered well enough (correlations above .98 between original and estimated means) only when at most one item out of five was affected by relatively large bias (and in no more than 50% of the groups) while the remaining items were affected by only tiny deviations from average item parameters (Pokropek, Davidov and Schmidt, 2019[111]). Furthermore, the alignment method will not lead to the estimation of the correct theoretical model: the estimated model is almost guaranteed to be "the wrong model" (it is likely to be over-parametrised in most situations). The alignment method encourages comparisons of item parameters across groups, when in many cases, the number of respondents per item and group (particularly when items are administered according to an incomplete design) is not sufficient to support precise estimates at the group level. Finally, the typical quadratic loss function used in the second optimisation step is sensitive to outliers; and the basic idea can be applied to a multiplicity of loss functions, each leading to a different solution (e.g. in a MGCFA model, should deviations in intercepts be penalised differently from deviations in slope parameters, given that they are not on the same scale?). While most users rely on a "black-box" implementation of the alignment method in the M*plus* software, there is still need for research on the decision rules and the properties of the invariance index in a variety of situations (sample size, number of items, number of response categories, number of groups, link functions, etc.).

Bayesian Approximate Invariance Methods

In situations in which perfect equivalence of measurements is understood to be an unrealistic ideal, a more elegant solution is to introduce greater realism in the models, e.g. by allowing all parameters to vary within a certain wiggle room. In such "approximate invariance" models, measurement parameters can vary across groups, according to a certain distribution (e.g. a normal distribution with a common mean and variance for the measurement parameter). Bayesian estimation is needed in such situations to make the problem computationally tractable.

The application of Bayesian random parameter models to measurement invariance situations was first proposed in the IRT framework (De Jong, Steenkamp and Fox, 2007[45]), then extended to MGCFA (Bayesian Structural Equation Modelling) (van de Schoot et al., 2013[30]; Muthén and Asparouhov, 2018[21]). Bayesian estimation of Approximate Measurement Invariance (AMI) models usually starts with informative priors, such as knowledge that differences in model parameters across groups are usually "small", and updates these priors with the information contained in the data.

In typical applications of Bayesian-AMI, priors loom quite large on the final solution. Indeed, the typical sample sizes per group and item imply significant uncertainty for the estimates of group-specific random deviations; and the number of groups is rarely large enough to provide significant information on the distribution of these random deviations from common parameters. On the other hand, in situations with many parameters and large samples, convergence in these models is hard to achieve, with a single model often running for several days before converging to a solution.

Rens van de Schoot suggested that because of the dependence on priors, practitioners should conduct a sensitivity analysis before drawing substantive conclusions; i.e. estimate models with different priors and verify the robustness of the resulting claims (see Chapter 3). In general, there was no consensus on how to rank models based on different priors (and thus, select the "best" priors and models): Jean-Paul Fox highlighted that criteria such as posterior predictive p-values (PPP) or deviance information criteria (DIC) should not be used to compare models with the same number of parameters. On the other hand, using the same priors for all parameters may be just as unrealistic as assuming that there is no variation in measurement parameters, but tailored priors may invite an abuse of "researcher degrees of freedom" (Simmons, Nelson and Simonsohn, 2011[113]), especially if they influence the conclusions strongly.

All presenters and discussants also highlighted the risk presented by "outlier" groups, which may "pull" the estimates of the parameter means and introduce bias in comparisons of latent means. This risk was well-illustrated in the simulation study presented by Arthur Pokropek: fitting an "approximate invariance model" to situations where a few groups and items are affected by large bias (partial invariance) leads to bias in the estimation of latent means (Pokropek, Davidov and Schmidt, 2019[111]). Another undesirable property of these methods is that the "ideal" situation in which there is no variation in measurement parameters is, now, a limit case, and a "corner solution" for the estimation procedure.

In response to some of these shortcomings, Jean-Paul Fox presented an alternative approach to assess whether the data support full invariance or only approximate invariance of measurements, which he illustrated in the IRT case (see Chapter 4). The approach, which was recently presented in Fox, Mulder and Sinharay (2017[42]), is based on the intuition that the marginal model obtained by integrating out the random parameters from a one-parameter IRT model is simply a fixed-effect model with a particular structure for the covariance of residuals. Therefore it is possible to conduct an analysis of residuals from the simpler model to identify (using Bayes Factor tests) whether a complex covariance structure (indicating AMI) fits the data better than a simple covariance structure (indicating full invariance), without the need to specify proper priors. Several discussants highlighted merits with this approach – including its simplicity, and the limited computational resources required. The approach is being further developed to a more general class of models.

A common problem with current Bayesian approaches for measurement invariance is that they still cannot handle complex survey design (weights, stratification, clustering) easily.

Complex random parameter models also have identification issues, which lead to convergence issues. When interest lies in identifying the sources of measurement non-invariance (such as the most problematic groups and items), some post-estimation diagnostic methods have been proposed, but their validity and reliability remains to be confirmed in simulation studies. On the other hand, when certain known features (such as writing system, level of development, climate zone …) are expected to interfere with measurements in some predictable ways, this information can be incorporated in the priors used to estimate Bayesian random parameter models.

General discussion

Throughout the discussion, several participants observed how the distinction between (MG)CFA and (MG)IRT worlds is largely artificial. Many recent developments in the field of measurement invariance seem to come from "rediscovering" some of the tools of IRT in the CFA framework, and vice-versa; and much more can still be gained from more opportunities for the two communities of scholars and practitioners to meet and work together. For example, in situations where the objective is to compare scale means across groups, it may seem preferable to summarise the uncertainty affecting such comparisons in a "scale uncertainty" parameter, instead of presenting several comparisons derived under different assumptions, and risk confusion and scepticism among readers. The similarity between "measurement invariance" and "test linking" problems, would suggest the use of "link errors" in comparisons of scales across groups (OECD, 2017, pp. 176-179[114]; Robitzsch and Lüdtke, 2018[115]).

The recent developments in the field of measurement invariance research originated from the availability of greater computing power to deal with complex models, large sample sizes and the global reach of large-scale surveys. The application and simulation studies presented at the conference also repeatedly highlighted the importance of avoiding short scales (made of only 3 or 4 items) in situations of imperfect equivalence (and particularly, when large biases could affect some item/group pairs).

The discussion also highlighted a consensus among all participants that any procedure to address the possible violation of (full) measurement invariance must consider the non-comparability of scales as a possibility. A procedure that is blind to serious violations of measurement equivalence, and promises to turn any measurement into a comparable one, is just as useless as one that is overly sensitive to small, inconsequential violations of an ideal model of invariance.

Dealing with imperfect comparability of measurements when scaling and reporting categorical latent variables

In the second day of the conference, a short session was devoted to how latent class analysis (LCA) could deal with issues of non-invariance of measurements, as they arise in large-scale international surveys. The generic definition of invariance as a conditional independence property of the measurement model does also apply to latent class models; it implies that, conditional on (latent) class membership, response probabilities for the observed categories do not depend on group (e.g. country) membership. In generic latent class models where the classes are treated as nominal the different levels of invariance (configural, metric, scalar, …) do not have a clear equivalent; in contrast, different levels of invariance can be defined for latent class models in which classes are ordered (Kankaraš, Vermunt and Moors, 2011[116]).

The two presenters in this session – Michael Eid and Jeroen Vermunt – shared their experience and advice about conducting LCA on large-scale, international surveys with the audience.

A simple strategy to conduct LCA on international data sets is described by Eid and Diener (2001[98]) and by Kankaraš, Moors and Vermunt (2018[82]). It starts by fitting country-specific latent class models in exploratory mode to find the number of classes that are supported in each country; results are reviewed to check if all or some of the classes reflect similar patterns in responses across multiple countries.

In a second step, samples are pooled. If the number of classes found in the first step does not differ across countries, the assumption of full measurement invariance is tested by comparing the fit of the model without measurement invariance (i.e. allowing observed responses to reflect both class and country membership) and the model with measurement invariance. If the model with full measurement invariance does not fit the data, different forms of partial measurement invariance can be tested (e.g. only some classes or some items are measurement invariant). If the number of the classes differ between countries, it can be tested whether the classes that are present in all countries are measurement invariant or not. Models can be compared with likelihood ratio tests or on information criteria. This strategy however is very cumbersome to apply for more than a handful of countries and items (the number of classes tends to increase with the number of items), because of the large number of models to estimate and of country/class combinations to review.

A second strategy is better suited for international surveys with dozens of countries (see Chapter 5). In this strategy, the exploratory step to determine the optimal number of classes is conducted directly on the pooled dataset, assuming, in a first step, that only class membership (and not country membership) determines the response patterns, while group membership only influences the size of classes. An inspection of the results can provide useful information about whether the latent classes are present in all countries. Measurement non-equivalence can manifest itself, for example, by some classes that are only present in some countries (size equal to 0). If this or other reasons (such as translation issues, differential social desirability contexts,…) lead practitioners to suspect measurement non-equivalence, a model that allows responses to reflect not only class but also group membership would have to be specified; the more general model, however, would have a very large number of parameters and, if group size is small, result in unstable parameter estimates. A possible solution to both issues is to specify a multilevel latent class model, where countries themselves are conceptualised as the expression of some latent set. While it may appear artificial to apply multilevel modelling to countries (which are not randomly selected groups from some overarching population, in direct violation of one of the model's assumption), treating countries as a random factor can reveal interesting sets of countries which share a common culture or institutional setting, which manifests itself in survey responses.

There is still little methodological research about this second strategy. It was illustrated in practice by Michael Eid with an application on the TALIS dataset, which revealed several issues that practitioners may encounter:

- Conducting LCA in exploratory mode on large data sets can be very time consuming;
- Proper identification of latent classes and conditional response probabilities requires large samples (both overall, and at the group level in multigroup LCA);

- Relying only on statistical criteria, such as fit indices, does not always provide conclusive guidance regarding model selection, which must also be informed by priors and by qualitative judgements informed by the solution;
- Convergence issues are quite frequent with complex LCA models; default starting values may not be sufficient, and in this application, the search for the optimal number of level-2 classes (country sets) was interrupted because of convergence issues when more than 6 classes (for 38 countries) were specified for the solution.

As Jeroen Vermunt made clear in his presentation, the application of LCA to surveys potentially affected by in-equivalence can benefit from recent methodological developments, such as local fit indices for multilevel LCA (Nagelkerke, Oberski and Vermunt, 2016[117]), or the extension of tools for quantifying the substantive impact of violations of invariance assumptions to categorical latent variables (Oberski, Vermunt and Moors, 2015[118]). Multilevel LCA also falls within a more general class of multilevel mixture models, which are a way of dealing with heterogeneity by modelling the responses as reflecting different measurement models, with each model specific to a latent class of individuals or groups (e.g. countries).

Improving the design of questionnaires for greater comparability of responses

The conference concluded with a discussion and demonstration of several innovative item types intended to reduce the incidence of measurement non-equivalence in cross-cultural research. Jonas Bertling presented how anchoring vignettes and situational judgement tests (SJT) were used in PISA 2012 to complement more traditional frequency and agreement (Likert) scales (Kyllonen and Bertling, 2013[119]). Pauline Slot and Trude Nilsen showed how situational judgement tests are being used in the TALIS Starting Strong Survey (TALIS-3S), aimed at pre-school educators. These illustrations highlighted the rationale for using these item types, and the practical choices that need to be made when analysing the responses and reporting them on a scale. Jia He and Patrick Kyllonen then introduced a more general discussion on these approaches.

The discussion highlighted some strong reservations about the use of anchoring vignettes, a method proposed by King et al. (2004[120]), whereby the subjective nature of response scale is overcome by asking respondents to report not only a self-assessment on the scale, but also, on the same scale, how they would assess several hypothetical individuals, presented in short vignettes. In practice, the ratings observed for the hypothetical individuals often violate the expected ratings, particularly in low-ability groups (perhaps due to respondent disengagement, or to the high cognitive load that this procedure imposes to participants). And some of the purported "improvements" in reliability and validity may be artificial, simply due to mathematical properties of the method, rather than to the substantive information gained about the response style of individuals (von Davier et al., 2018[121]). Furthermore, the choice of the vignettes appears not to be neutral with respect to the substantive conclusions (Stankov, Lee and von Davier, 2018[122]).

More optimism was expressed about the potential of SJTs (Lievens, Peeters and Schollaert, 2008[123]; McDaniel et al., 2007[124]). Situational judgement items present respondents with hypothetical situations, and ask them to report "how likely" they are to act in certain ways; answers may be provided as "behavioural tendencies ("very likely", etc.) or as forced choice (e.g. by selecting the "most likely" and "least likely" option). The hypothetical nature of the situation seems to reduce some social desirability biases, which often affect behavioural self-reports.

Questionnaire developers should nevertheless keep in mind that SJTs may not lend themselves to all sorts of constructs, are relatively long to administer, and may be more appropriate for high-ability populations, such as teachers, due to the high cognitive load of thinking through hypothetical scenarios, particularly when administered in written form. Situational judgement test items should only be administered to populations for which familiarity with the described situation can be assumed.

In the end, it appeared clear that significant gains in comparability of survey responses across groups of respondents can also be made by following simple and universal design principles, which are not always met in practice: write items that are clearer, more concrete, behavioural, simple, less abstract. Short scales (of only 3-4 items) should not be used in situations where non-equivalence at the item-level is a possibility. This may well mean that, rather than including many constructs, future surveys should include fewer, but better and longer scales.

Despite the limitations in these innovative item types, and the reservations expressed about anchoring vignettes, all discussants and participants agreed that greater variety in response formats may be desirable to triangulate findings and ensure they are not driven by surface features of the instruments. For example, it may be desirable to measure a certain construct through both forced choice items and Likert items.

References

Andrews, R. and I. Currim (2003), "A Comparison of Segment Retention Criteria for Finite Mixture Logit Models", *Journal of Marketing Research*, Vol. 40/2, pp. 235-243, http://dx.doi.org/10.1509/jmkr.40.2.235.19225. [94]

Asparouhov, T. and B. Muthén (2014), "Multiple-Group Factor Analysis Alignment", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 21/4, pp. 495-508, http://dx.doi.org/10.1080/10705511.2014.919210. [19]

Asparouhov, T. and B. Muthén (2009), "Exploratory Structural Equation Modeling", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 16/3, pp. 397-438, http://dx.doi.org/10.1080/10705510903008204. [7]

Azevedo, C., D. Andrade and J. Fox (2012), "A Bayesian generalized multiple group IRT model with model-fit assessment tools", *Computational Statistics & Data Analysis*, Vol. 56/12, pp. 4399-4412, http://dx.doi.org/10.1016/j.csda.2012.03.017. [60]

Bartholomew, D., M. Knott and I. Moustaki (2011), *Latent variable models and factor analysis : a unified approach.*, Wiley, https://www.wiley.com/en-us/Latent+Variable+Models+and+Factor+Analysis%3A+A+Unified+Approach%2C+3rd+Edition-p-9780470971925 (accessed on 3 January 2019). [71]

Bock, R. and M. Zimowski (1997), "Multiple Group IRT", in *Handbook of Modern Item Response Theory*, Springer New York, New York, NY, http://dx.doi.org/10.1007/978-1-4757-2691-6_25. [59]

Boer, D., K. Hanke and J. He (2018), "On Detecting Systematic Measurement Error in Cross-Cultural Research: A Review and Critical Reflection on Equivalence and Invariance Tests", *Journal of Cross-Cultural Psychology*, Vol. 49/5, pp. 713-734, http://dx.doi.org/10.1177/0022022117749042. [108]

Boyce, J. and A. Bowers (2016), "Principal Turnover: Are There Different Types of Principals Who Move From or Leave Their Schools? A Latent Class Analysis of the 2007–2008 Schools and Staffing Survey and the 2008–2009 Principal Follow-Up Survey", *Leadership and Policy in Schools*, Vol. 15/3, pp. 237-272, http://dx.doi.org/10.1080/15700763.2015.1047033. [79]

Buchholz, J. and J. Hartig (2017), "Comparing Attitudes Across Groups: An IRT-Based Item-Fit Statistic for the Analysis of Measurement Invariance", *Applied Psychological Measurement*, p. 014662161774832, http://dx.doi.org/10.1177/0146621617748323. [110]

Byrne, B., R. Shavelson and B. Muthén (1989), "Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance.", *Psychological Bulletin*, Vol. 105/3, pp. 456-466, http://dx.doi.org/10.1037/0033-2909.105.3.456. [6]

Chen, F. (2007), "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 14/3, pp. 464-504, http://dx.doi.org/10.1080/10705510701301834. [15]

Cieciuch, J., E. Davidov and P. Schmidt (2018), "Alignment Optimization: Estimation of the Most Trustworthy Means in Cross-Cultural Studies Even in the Presence of Noninvariance", in Davidov, E., P. Schmidt and J. Billiet (eds.), *Cross-cultural analysis : methods and applications*, Routledge, http://dx.doi.org/10.4324/9781315537078. [22]

Clogg, C. (1995), "Latent Class Models", in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, Springer US, Boston, MA, http://dx.doi.org/10.1007/978-1-4899-1292-3_6. [67]

Clogg, C. and L. Goodman (1985), "Simultaneous Latent Structure Analysis in Several Groups", *Sociological Methodology*, Vol. 15, p. 81, http://dx.doi.org/10.2307/270847. [80]

Collins, L. and S. Lanza (2010), *Latent class and latent transition analysis. With applications in the social, behavioral, and health sciences*, Wiley. [68]

Davidov, E. et al. (2015), "The Comparability of Measurements of Attitudes toward Immigration in the European Social Survey", *Public Opinion Quarterly*, Vol. 79/S1, pp. 244-266, http://dx.doi.org/10.1093/poq/nfv008. [16]

Davidov, E. et al. (2014), "Measurement Equivalence in Cross-National Research", *Annual Review of Sociology*, Vol. 40/1, pp. 55-75, http://dx.doi.org/10.1146/annurev-soc-071913-043137. [9]

De Jong, M., J. Steenkamp and J. Fox (2007), "Relaxing Measurement Invariance in Cross-National Consumer Research Using a Hierarchical IRT Model", *Journal of Consumer Research*, Vol. 34/2, pp. 260-278, http://dx.doi.org/10.1086/518532. [45]

Dias, J. (2006), "Latent Class Analysis and Model Selection", in *From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization*, Springer-Verlag, Berlin/Heidelberg, http://dx.doi.org/10.1007/3-540-31314-1_10. [95]

Donoghue, J. and N. Allen (1993), "Thin Versus Thick Matching in the Mantel-Haenszel Procedure for Detecting DIF", *Journal of Educational Statistics*, Vol. 18/2, pp. 131-154, http://dx.doi.org/10.3102/10769986018002131. [48]

Dorans, N. and P. Holland (1993), "DIF detection and description: Mantel-Haenszel and standardization", in Holland, P. and H. Wainer (eds.), *Differential item functioning*, Erlbaum. [49]

Drossel, K. and B. Eickelmann (2017), "Teachers' participation in professional development concerning the implementation of new technologies in class: a latent class analysis of teachers and the relationship with the use of computers, ICT self-efficacy and emphasis on teaching ICT skills", *Large-scale Assessments in Education*, Vol. 5/1, http://dx.doi.org/10.1186/s40536-017-0053-7. [73]

Eid, M. and E. Diener (2001), "Norms for experiencing emotions in different cultures: Inter- and intranational differences.", *Journal of Personality and Social Psychology*, Vol. 81/5, pp. 869-885, http://dx.doi.org/10.1037/0022-3514.81.5.869. [98]

Eid, M., R. Langeheine and E. Diener (2003), "Comparing Typological Structures Across Cultures By Multigroup Latent Class Analysis", *Journal of Cross-Cultural Psychology*, Vol. 34/2, pp. 195-210, http://dx.doi.org/10.1177/0022022102250427. [81]

Fagginger Auer, M. et al. (2016), "Multilevel Latent Class Analysis for Large-Scale Educational Assessment Data: Exploring the Relation Between the Curriculum and Students' Mathematical Strategies", *Applied Measurement in Education*, Vol. 29/2, pp. 144-159, http://dx.doi.org/10.1080/08957347.2016.1138959. [76]

Fonseca, J. and M. Cardoso (2007), "Mixture-model cluster analysis using information theoretical criteria", *Intelligent Data Analysis*, Vol. 11/2, pp. 155-173, http://dx.doi.org/10.3233/ida-2007-11204. [96]

Fox, J. (2010), *Bayesian Item Response Modeling*, Springer New York, New York, NY, http://dx.doi.org/10.1007/978-1-4419-0742-4. [37]

Fox, J., J. Mulder and S. Sinharay (2017), "Bayes Factor Covariance Testing in Item Response Models", *Psychometrika*, Vol. 82/4, pp. 979-1006, http://dx.doi.org/10.1007/s11336-017-9577-6. [42]

Gelman, A. et al. (2003), *Bayesian data analysis*, Chapman & Hall. [57]

Gelman, A., X. Meng and H. Stern (1996), "Posterior predictive assessment of model fitness via realized discrepancies", *Statistica Sinica*, Vol. 6, pp. 733–807. [66]

Haberman, S. (2009), "Linking Parameter Estimates Derived From An Item Response Model Through Separate Calibrations", *ETS Research Report Series*, Vol. 2009/2, pp. i-9, http://dx.doi.org/10.1002/j.2333-8504.2009.tb02197.x. [112]

Hagenaars, J. and A. McCutcheon (eds.) (2002), *Applied Latent Class Analysis*, Cambridge University Press, Cambridge, http://dx.doi.org/10.1017/cbo9780511499531. [69]

Hallquist, M. and J. Wiley (2018), *MplusAutomation: An R Package for Facilitating Large-Scale Latent Variable Analyses in Mplus*, https://github.com/michaelhallquist/MplusAutomation (accessed on 4 January 2019). [34]

Hambleton, R. and H. Rogers (1989), "Detecting Potentially Biased Test Items: Comparison of IRT Area and Mantel-Haenszel Methods", *Applied Measurement in Education*, Vol. 2/4, pp. 313-334, http://dx.doi.org/10.1207/s15324818ame0204_4. [46]

Heinen, T. (1996), *Latent class and discrete latent trait models: Similarities and differences*, Sage. [72]

Hoijtink, H. (2012), *Informative hypotheses. Theory and practice for behavioral and social scientists*, Chapman & Hall/CRC. [54]

Hoijtink, H. and R. van de Schoot (2018), "Testing small variance priors using prior-posterior predictive p values.", *Psychological Methods*, Vol. 23/3, pp. 561-569, http://dx.doi.org/10.1037/met0000131. [35]

Holland, P. and H. Wainer (1993), *Differential Item Functioning*, Erlbaum, Hillsdale, NJ. [4]

Horn, J. and J. Mcardle (1992), "A practical and theoretical guide to measurement invariance in aging research", *Experimental Aging Research*, Vol. 18/3, pp. 117-144, http://dx.doi.org/10.1080/03610739208253916. [104]

Hu, L. and P. Bentler (1999), "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 6/1, pp. 1-55, http://dx.doi.org/10.1080/10705519909540118. [14]

Jeffreys, H. (1961), *Theory of probability.*, Oxford University Press. [53]

Jöreskog, K. (1971), "Simultaneous factor analysis in several populations", *Psychometrika*, Vol. 36/4, pp. 409-426, http://dx.doi.org/10.1007/bf02291366. [2]

Jöreskog, K. (1969), "A general approach to confirmatory maximum likelihood factor analysis", *Psychometrika*, Vol. 34/2, pp. 183-202, http://dx.doi.org/10.1007/bf02289343. [1]

Kankaraš, M., G. Moors and J. Vermunt (2018), "Testing for Measurement Invariance With Latent Class Analysis", in Davidov, E., P. Schmidt and J. Billiet (eds.), *Cross-cultural analysis : methods and applications*, Routledge, http://dx.doi.org/10.4324/9781315537078. [82]

Kankaraš, M., J. Vermunt and G. Moors (2011), "Measurement Equivalence of Ordinal Items: A Comparison of Factor Analytic, Item Response Theory, and Latent Class Approaches", *Sociological Methods & Research*, Vol. 40/2, pp. 279-310, http://dx.doi.org/10.1177/0049124111405301. [116]

Kass, R. and A. Raftery (1995), "Bayes Factors", *Journal of the American Statistical Association*, Vol. 90/430, pp. 773-795, http://dx.doi.org/10.1080/01621459.1995.10476572. [52]

Kelcey, B., D. McGinn and H. Hill (2014), "Approximate measurement invariance in cross-classified rater-mediated assessments", *Frontiers in Psychology*, Vol. 5, http://dx.doi.org/10.3389/fpsyg.2014.01469. [39]

Kim, E. et al. (2017), "Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 24/4, pp. 524-544, http://dx.doi.org/10.1080/10705511.2017.1304822. [44]

King, G. et al. (2004), "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research", *American Political Science Review*, Vol. 98/01, pp. 191-207, http://dx.doi.org/10.1017/S000305540400108X. [120]

Klugkist, I. and H. Hoijtink (2007), "The Bayes factor for inequality and about equality constrained models", *Computational Statistics & Data Analysis*, Vol. 51/12, pp. 6367-6379, http://dx.doi.org/10.1016/j.csda.2007.01.024. [43]

Kyllonen, P. and J. Bertling (2013), "Innovative Questionnaire Assessment Methods to Increase Cross-Country Comparability", in Rutkowski, L., M. von Davier and D. Rutkowski (eds.), *Handbook of International large-scale assessment : background, technical issues, and methods of data analysis*, Chapman and Hall/CRC. [119]

Lancaster, H. (1965), "The Helmert Matrices", *The American Mathematical Monthly*, Vol. 72/1, pp. 4-12, http://dx.doi.org/10.1080/00029890.1965.11970483. [62]

Langeheine, R. (1988), "New Developments in Latent Class Theory", in *Latent Trait and Latent Class Models*, Springer US, Boston, MA, http://dx.doi.org/10.1007/978-1-4757-5644-9_5. [90]

Langeheine, R., J. Pannekoek and F. Van De Pol (1996), "Bootstrapping Goodness-of-Fit Measures in Categorical Data Analysis", *Sociological Methods & Research*, Vol. 24/4, pp. 492-516, http://dx.doi.org/10.1177/0049124196024004004. [91]

Lazarsfeld, P. and N. Henry (1968), *Latent structure analysis*, Houghton Mifflin. [70]

Lee, H. and K. Geisinger (2015), "The Matching Criterion Purification for Differential Item Functioning Analyses in a Large-Scale Assessment", *Educational and Psychological Measurement*, Vol. 76/1, pp. 141-163, http://dx.doi.org/10.1177/0013164415585166. [51]

Lek, K. et al. (2018), "Approximate Measurement Invariance", in Johnson, Timothy P. Pennell, Beth-Ellen Stoop, Ineke A. L. Dorer, B. (ed.), *Advances in comparative survey methods : multinational, multiregional, and multicultural contexts (3MC)*, Wiley. [31]

Levy, R., R. Mislevy and S. Sinharay (2009), "Posterior Predictive Model Checking for Multidimensionality in Item Response Theory", *Applied Psychological Measurement*, Vol. 33/7, pp. 519-537, http://dx.doi.org/10.1177/0146621608329504. [58]

Lievens, F., H. Peeters and E. Schollaert (2008), "Situational judgment tests: a review of recent research", *Personnel Review*, Vol. 37/4, pp. 426-441, http://dx.doi.org/10.1108/00483480810877598. [123]

Lin, S. and W. Tai (2015), "Latent Class Analysis of Students' Mathematics Learning Strategies and the Relationship between Learning Strategy and Mathematical Literacy", *Universal Journal of Educational Research*, Vol. 3/6, pp. 390-395, http://dx.doi.org/10.13189/ujer.2015.030606. [78]

Lommen, M., R. van de Schoot and I. Engelhard (2014), "The experience of traumatic events disrupts the measurement invariance of a posttraumatic stress scale", *Frontiers in Psychology*, Vol. 5, p. 1304, http://dx.doi.org/10.3389/fpsyg.2014.01304. [102]

Lukočienė, O., R. Varriale and J. Vermunt (2010), "6. The Simultaneous Decision(s) about the Number of Lower- and Higher-Level Classes in Multilevel Latent Class Analysis", *Sociological Methodology*, Vol. 40/1, pp. 247-283, http://dx.doi.org/10.1111/j.1467-9531.2010.01231.x. [99]

Magis, D., S. Béland and G. Raiche (2015), *difR: Collection of Methods to Detect Dichotomous Differential Item Functioning (DIF)*, https://CRAN.R-project.org/package=difR. [47]

McDaniel, M. et al. (2007), "Situational Judgment Tests, Response Instructions, And Validity: A Meta-Analysis", *Personnel Psychology*, Vol. 60/1, pp. 63-91, http://dx.doi.org/10.1111/j.1744-6570.2007.00065.x. [124]

Mellenbergh, G. (1989), "Item bias and item response theory", *International Journal of Educational Research*, Vol. 13/2, pp. 127-143, http://dx.doi.org/10.1016/0883-0355(89)90002-5. [103]

Meredith, W. (1993), "Measurement invariance, factor analysis and factorial invariance", *Psychometrika*, Vol. 58/4, pp. 525-543, http://dx.doi.org/10.1007/BF02294825. [8]

Meuleman, B. (2012), "When are item intercept differences substantively relevant in measurement invariance testing?", in *Methods, Theories, and Empirical Applications in the Social Sciences*, VS Verlag für Sozialwissenschaften, Wiesbaden, http://dx.doi.org/10.1007/978-3-531-18898-0_13. [17]

Meuleman, B. and J. Billiet (2012), "Measuring Attitudes toward Immigration in Europe: The Cross-Cultural Validity of the ESS Immigration Scales", *Ask: Research and Methods*, Vol. 21/1, pp. 5-29. [32]

Millsap, R. (2011), *Statistical approaches to measurement invariance*, Taylor & Francis Group. [26]

Munck, I., C. Barber and J. Torney-Purta (2017), "Measurement Invariance in Comparing Attitudes Toward Immigrants Among Youth Across Europe in 1999 and 2009", *Sociological Methods & Research*, Vol. 47/4, pp. 687-728, http://dx.doi.org/10.1177/0049124117729691. [23]

Muthén, B. and T. Asparouhov (2018), "Recent Methods for the Study of Measurement Invariance With Many Groups", *Sociological Methods & Research*, Vol. 47/4, pp. 637-664, http://dx.doi.org/10.1177/0049124117701488. [21]

Muthén, B. and T. Asparouhov (2014), "IRT studies of many groups: the alignment method", *Frontiers in Psychology*, Vol. 5, p. 978, http://dx.doi.org/10.3389/fpsyg.2014.00978. [20]

Muthén, B. and T. Asparouhov (2013), "BSEM Measurement Invariance Analysis. Mplus Web Notes: No. 17", *Mplus Web Notes*, No. 17, Mplus, http://www.statmodel.com/examples/webnotes/webnote17.pdf (accessed on 4 January 2019). [29]

Muthén, B. and T. Asparouhov (2013), "New methods for the study of measurement invariance with many groups", Mplus, http://statmodel2.com/download/PolAn.pdf (accessed on 4 January 2019). [27]

Muthén, B. and T. Asparouhov (2012), "Bayesian structural equation modeling: A more flexible representation of substantive theory.", *Psychological Methods*, Vol. 17/3, pp. 313-335, http://dx.doi.org/10.1037/a0026802. [28]

Muthén, L. and B. Muthén (1998-2017), *Mplus User's Guide*, Muthén & Muthén. [25]

Nagelkerke, E., D. Oberski and J. Vermunt (2016), "Goodness-of-fit of Multilevel Latent Class Models for Categorical Data", *Sociological Methodology*, Vol. 46/1, pp. 252-282, http://dx.doi.org/10.1177/0081175015581379. [117]

Oberski, D. (2014), "Evaluating Sensitivity of Parameters of Interest to Measurement Invariance in Latent Variable Models", *Political Analysis*, Vol. 22/01, pp. 45-60, http://dx.doi.org/10.1093/pan/mpt014. [18]

Oberski, D., J. Vermunt and G. Moors (2015), "Evaluating Measurement Invariance in Categorical Data Latent Variable Models with the EPC-Interest", *Political Analysis*, Vol. 23/04, pp. 550-563, http://dx.doi.org/10.1093/pan/mpv020. [118]

OECD (2017), *PISA 2015 Technical Report*, OECD publishing, http://www.oecd.org/pisa/data/2015-technical-report/ (accessed on 27 November 2017). [114]

OECD (2014), *TALIS 2013 Technical Report*, OECD Publishing, http://www.oecd.org/education/school/TALIS-technical-report-2013.pdf. [83]

O'Hagan, A. (1995), "Fractional Bayes factors for model comparison", *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57/1, pp. 99-138. [55]

Oliveri, M. et al. (2014), "Uncovering Substantive Patterns in Student Responses in International Large-Scale Assessments—Comparing a Latent Class to a Manifest DIF Approach", *International Journal of Testing*, Vol. 14/3, pp. 265-287, http://dx.doi.org/10.1080/15305058.2014.891223. [75]

Plummer, M. et al. (2006), "CODA: Convergence diagnosis and output analysis for MCMC", *R news*, Vol. 6/1, pp. 7-11. [64]

Pokropek, A., E. Davidov and P. Schmidt (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*, pp. 1-21, http://dx.doi.org/10.1080/10705511.2018.1561293. [111]

Poznyak, D. et al. (2013), "Trust in American Government: Longitudinal Measurement Equivalence in the ANES, 1964–2008", *Social Indicators Research*, Vol. 118/2, pp. 741-758, http://dx.doi.org/10.1007/s11205-013-0441-5. [24]

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org (accessed on 4 January 2019). [33]

R Core Team (2014), *R: A language and environment for statistical computing*, http://www.r-project.org/. [63]

Rabe-Hesketh, S. and A. Skrondal (2006), "Multilevel modelling of complex survey data", *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 169/4, pp. 805-827, http://dx.doi.org/10.1111/j.1467-985X.2006.00426.x. [61]

Raftery, A. (1995), "Bayesian Model Selection in Social Research", *Sociological Methodology*, Vol. 25, p. 111, http://dx.doi.org/10.2307/271063. [56]

Rasch, G. (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute for Educational Research, Copenhagen, Denmark. [3]

Robitzsch, A. and O. Lüdtke (2018), "Linking errors in international large-scale assessments: calculation of standard errors for trend estimation", *Assessment in Education: Principles, Policy & Practice*, pp. 1-22, http://dx.doi.org/10.1080/0969594X.2018.1433633. [115]

Rutkowski, L. and D. Svetina (2016), "Measurement Invariance in International Surveys: Categorical Indicators and Fit Measure Performance", *Applied Measurement in Education*, Vol. 30/1, pp. 39-51, http://dx.doi.org/10.1080/08957347.2016.1243540. [107]

Rutkowski, L. and D. Svetina (2013), "Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys", *Educational and Psychological Measurement*, Vol. 74/1, pp. 31-57, http://dx.doi.org/10.1177/0013164413498257. [106]

Saris, W., A. Satorra and W. van der Veld (2009), "Testing Structural Equation Models or Detection of Misspecifications?", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 16/4, pp. 561-582, http://dx.doi.org/10.1080/10705510903203433. [12]

Schafer, J. and J. Kang (2013), *LCCA package for R Users' Guide (Version 1.1.0)*, The Methodology Center, Penn State, https://methodology.psu.edu/node/1047. [100]

Simmons, J., L. Nelson and U. Simonsohn (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant", *Psychological Science*, Vol. 22/11, pp. 1359-1366, http://dx.doi.org/10.1177/0956797611417632. [113]

Sinharay, S., M. Johnson and H. Stern (2006), "Posterior Predictive Assessment of Item Response Theory Models", *Applied Psychological Measurement*, Vol. 30/4, pp. 298-321, http://dx.doi.org/10.1177/0146621605285517. [65]

Stankov, L., J. Lee and M. von Davier (2018), "A Note on Construct Validity of the Anchoring Method in PISA 2012", *Journal of Psychoeducational Assessment*, Vol. 36/7, pp. 709-724, http://dx.doi.org/10.1177/0734282917702270. [122]

Steenkamp, J. and H. Baumgartner (1998), "Assessing Measurement Invariance in Cross-National Consumer Research", *Journal of Consumer Research*, Vol. 25/1, pp. 78-107, http://dx.doi.org/10.1086/209528. [10]

Tekle, F., D. Gudicha and J. Vermunt (2016), "Power analysis for the bootstrap likelihood ratio test for the number of classes in latent class models", *Advances in Data Analysis and Classification*, Vol. 10/2, pp. 209-224, http://dx.doi.org/10.1007/s11634-016-0251-0. [92]

Thissen, D., L. Steinberg and M. Gerrard (1986), "Beyond group-mean differences: The concept of item bias.", *Psychological Bulletin*, Vol. 99/1, pp. 118-128, http://dx.doi.org/10.1037/0033-2909.99.1.118. [36]

Thissen, D., L. Steinberg and H. Wainer (1993), "Detection of differential item functioning using the parameters of item response models", in Holland, P. and H. Wainer (eds.), *Differential item functioning*, Erlbaum. [50]

van de Schoot, R. et al. (2013), "Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance", *Frontiers in Psychology*, Vol. 4, p. 770, http://dx.doi.org/10.3389/fpsyg.2013.00770. [30]

van de Schoot, R., P. Lugtig and J. Hox (2012), "A checklist for testing measurement invariance", *European Journal of Developmental Psychology*, Vol. 9/4, pp. 486-492, http://dx.doi.org/10.1080/17405629.2012.686740. [105]

Van de Vijver, F. and K. Leung (1997), *Methods and Data Analysis for Cross-Cultural Research*, Sage, Newbury Park, CA. [5]

van de Vijver, F. and N. Tanzer (2004), "Bias and equivalence in cross-cultural assessment: an overview", *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, Vol. 54/2, pp. 119-135, http://dx.doi.org/10.1016/j.erap.2003.12.004. [38]

Vandenberg, R. and C. Lance (2000), "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research", *Organizational Research Methods*, Vol. 3/1, pp. 4-70, http://dx.doi.org/10.1177/109442810031002. [11]

Verhagen, A. and J. Fox (2013), "Bayesian tests of measurement invariance", *British Journal of Mathematical and Statistical Psychology*, Vol. 66/3, pp. 383-401, http://dx.doi.org/10.1111/j.2044-8317.2012.02059.x. [40]

Verhagen, J. et al. (2016), "Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models", *Journal of Mathematical Psychology*, Vol. 72, pp. 171-182, http://dx.doi.org/10.1016/j.jmp.2015.06.005. [41]

Vermunt, J. (2008), "Latent class and finite mixture models for multilevel data sets", *Statistical Methods in Medical Research*, Vol. 17/1, pp. 33-51, http://dx.doi.org/10.1177/0962280207081238. [85]

Vermunt, J. (2003), "7. Multilevel Latent Class Models", *Sociological Methodology*, Vol. 33/1, pp. 213-239, http://dx.doi.org/10.1111/j.0081-1750.2003.t01-1-00131.x. [84]

Vermunt, J. (1997), *LEM 1.0: A general program for the analysis of categorical data*, Tilburg University. [101]

Vermunt, J. and J. Magidson (2016), *Technical guide for Latent GOLD 5.1: Basic, advanced, and syntax*, Statistical Innovations. [86]

Vermunt, J. and J. Magidson (2005), *Latent Gold 4.0 user's guide*, Statistical Innovations. [97]

von Davier, M. (2010), "Hierarchical mixtures of diagnostic models", *Psychological Test and Assessment Modeling*, Vol. 52/10, pp. 8-28, http://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2010/02_vonDavier.pdf (accessed on 3 January 2019). [88]

von Davier, M. (2005), *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models*, ETS. [87]

von Davier, M. and J. Rost (2016), "Logistic Mixture-Distribution Response Models", in van der Linden, W. (ed.), *Handbook of Item Response Theory, Volume One*, Chapman and Hall/CRC, http://dx.doi.org/10.1201/9781315374512. [89]

von Davier, M. et al. (2018), "The Effects of Vignette Scoring on Reliability and Validity of Self-Reports", *Applied Psychological Measurement*, Vol. 42/4, pp. 291-306, http://dx.doi.org/10.1177/0146621617730389. [121]

von Davier, M. et al. (2018), "Evaluating Item Response Theory Linking and Model Fit for Data from PISA 2000–2012", *Assessment in Education: Principles, Policy & Practice*, Vol. Special Issue. [109]

West, S., A. Taylor and W. Wu (2012), "Model fit and model selection in structural equation modeling", in Hoyle, R. (ed.), *Handbook of structural equation modeling*, Guilford Press. [13]

Yalcin, S. (2017), "Determining the Relationships between Selected Variables and Latent Classes in Students' PISA Achievement", *International Journal of Research in Education and Science*, pp. 589-589, http://dx.doi.org/10.21890/ijres.328089. [77]

Yang, C. and C. Yang (2007), "Separating Latent Classes by Information Criteria", *Journal of Classification*, Vol. 24/2, pp. 183-203, http://dx.doi.org/10.1007/s00357-007-0010-1. [93]

Zhang, Y., R. Watermann and A. Daniel (2016), "Are multiple goals in elementary students beneficial for their school achievement? A latent class analysis", *Learning and Individual Differences*, Vol. 51, pp. 100-110, http://dx.doi.org/10.1016/j.lindif.2016.08.023. [74]