OECD Education Working Papers No. 124

Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013

Jia He,

Katarzyna Kubacka

https://dx.doi.org/10.1787/5jrp6fwtmhf2-en

OECD

Unclassified **EDU/WKP(2015)13**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **03-Dec-2015**

English - Or. English

DIRECTORATE FOR EDUCATION AND SKILLS

EDU/WKP(2015)13
Unclassified

Cancels & replaces the same document of 23 November 2015

DATA COMPARABILITY IN THE TEACHING AND LEARNING INTERNATIONAL SURVEY (TALIS) 2008 AND 2013

By Jia He and Katarzyna Kubacka

OECD Education Working Paper no. 124

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

Contact:
Jia He, German Institute for International Educational Research, (jia.he@dipf.de)
Katarzyna Kubacka, Analyst, Directorate for Education and Skills, (katarzyna.kubacka@oecd.org)

JT03387693

Complete document available on OLIS in its original format

English - Or. English

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

The authors would like to thank Julie Bélanger (RAND Europe, formerly OECD Directorate for Education and Skills), Noémie Le Donné and Karine Tremblay (OECD Directorate for Education and Skills), and Fons van de Vijver (Tilburg University) for their helpful comments on the research proposal underlying this paper and earlier versions of the draft; and to Emily Groves, Camilla Lorentzen and Jennifer Cannon (OECD Directorate for Education and Skills) for their help in finalising the paper.

ABSTRACT

This report focuses on data comparability of scale scores in the Teaching and Learning International Survey (TALIS).

Valid cross-cultural comparisons of TALIS data are vital in providing input for evidence-based policy making and in promoting the equity and effectiveness of teacher policies. For this purpose, an investigation of data comparability is a prerequisite for any meaningful cross-cultural comparison.

TALIS involves a large number of countries and economies, and has used rather strict conventional statistical methods to test comparability. Thus, many scales in TALIS do not reach the level of comparability that allows direct comparisons of scale scores. To facilitate the effective data analysis of TALIS and maximise its policy implications, this project: (1) uses a more flexible statistical method to test comparability, and (2) investigates the level and sources of scale data incomparability.

With teacher and principal self-report data from the two rounds of TALIS (2008 and 2013), three studies are carried out to address these issues. Study 1 compares the conventional statistical method with more flexible Bayesian approximate invariance testing in scale data comparability testing. Study 2 investigates whether scale characteristics (e.g. scale length, item length, number of response options, and self-evaluative components) are associated with data comparability in principal and teacher scales. Finally, Study 3 examines the specific cultural variations that contribute to the lack of comparability. It tests the comparability of the *Satisfaction with Current Work Environment* scale (a key outcome construct in TALIS) between each participating country or economy with a pooled international average reference group.

The paper concludes with a discussion of the implications for large-scale survey design and data analyses such as using more flexible psychometric method to test comparability and using fewer response options in items forming scales.

RÉSUMÉ

Ce rapport étudie la comparabilité des données relatives aux scores obtenus sur les différentes échelles de l'Enquête internationale sur l'enseignement et l'apprentissage (TALIS).

Des comparaisons fiables des données de TALIS entre les différentes cultures sont vitales pour contribuer à l'élaboration de politiques fondées sur des informations probantes ainsi que pour promouvoir l'équité et l'efficacité des politiques concernant les enseignants. Dans cette optique, une étude de la comparabilité des données est indispensable pour permettre d'établir des comparaisons pertinentes entre les différentes cultures.

Dans la mesure où l'enquête TALIS implique de nombreux pays et économies, des méthodes statistiques conventionnelles assez strictes ont été adoptées pour tester la comparabilité des échelles construites. Ainsi, de nombreuses échelles de TALIS n'atteignent pas un niveau de comparabilité suffisant pour permettre des comparaisons directes des scores obtenus. Afin de favoriser une analyse efficace des données de TALIS et d'optimiser ses implications politiques, ce projet : (1) s'appuie sur une méthode statistique plus souple pour tester la comparabilité, et (2) étudie le niveau et les sources de non comparabilité des données des échelles.

Sur la base des données recueillies auprès des enseignants et des chefs d'établissement dans le cadre des deux cycles de TALIS (2008 et 2013), trois études sont réalisées pour examiner ces problématiques. L'étude 1 compare la méthode statistique conventionnelle avec l'approche plus souple du test bayésien de l'invariance approximative pour l'évaluation de la comparabilité des données des échelles. L'étude 2 examine si les caractéristiques des échelles (par ex., la longueur de l'échelle, la longueur des items, le nombre de modalités de réponse et les composantes d'auto-évaluation sont associées à la comparabilité des données échelles relatives aux chefs d'établissement et aux enseignants. Enfin, l'étude 3 examine les variations culturelles qui contribuent spécifiquement au manque de comparabilité. Elle teste la comparabilité de l'échelle de Satisfaction à l'égard de l'environnement de travail actuel (un construct majeur dans TALIS) entre chacun des pays ou des économies participants avec un groupe moyen international de référence.

Le document se conclut par un examen des implications pour la conception et l'analyse des données d'enquêtes à grande échelle, comme le recours à une méthode psychométrique plus souple pour tester la comparabilité et l'utilisation d'un nombre restreint de modalités de réponse pour les items qui constituent les échelles.

TABLE OF CONTENTS

Tables

Figures

Boxes

DATA COMPARABILITY IN THE TEACHING AND LEARNING INTERNATIONAL SURVEY (TALIS) 2008 AND 2013[1]

1. Introduction

Cross-country comparisons of large-scale international survey data, such as the Teaching and Learning International Survey (TALIS), are used to inform evidence-based policy making.

Box 1. What is TALIS?

TALIS is the first international survey programme to focus on the learning environment and the working conditions of teachers in schools. It asks teachers and school principals about their attitudes and experiences with their work, schools and classrooms. The first TALIS cycle – TALIS 2008 – focused on lower secondary education teachers and their principals. It sampled 200 schools in each of 24 countries and 20 teachers in each school. The second cycle – TALIS 2013 – similarly surveyed teachers and principals in 38 countries and economies. Thirty-four of these countries and economies carried out the survey in 2014, and four additional countries and economies collected data one year later, but using the same method and standards as other countries and were, therefore, added to the database.

The cross-country analyses of TALIS data help countries identify others facing similar challenges and learn about their policies. For instance, through linking deep forms of collaborative behaviours in teachers and their use of certain active teaching practices, it is found that some countries are particularly good at this, thus policy makers in other countries could follow successful examples and take measures to improve the education system in their countries (e.g. OECD, 2014a).

As results from cross-country comparisons of international surveys such as TALIS are used for policy making, valid comparisons are vital in promoting equity and ensuring the effectiveness of teacher policies (Meredith, 1993). Previously, the comparability of data was assumed; this implied that an equal score on a global measure means the same level of trait across countries. In contrast, new developments in cross-cultural research methods in the past two decades urge that data comparability should be conceptually and statistically demonstrated prior to any international comparison (e.g. Harkness, van de Vijver and Mohler, 2003; van de Vijver and Leung, 1997). Extant research suggests that without the demonstration of comparability, comparisons of cross-cultural data are at best ambiguous and at worst erroneous (Steenkamp and Baumgartner, 1998; Chen, 2008).

TALIS has been designed so that it utilises single-item measures and scales to measure teachers' and principals' experiences and attitudes. There are a number of advantages of using scales over single-item measures. For instance, multiple items in one scale can capture different facets of a concept and subsequently afford higher reliability and predictive validity than a single item (e.g. Diamantopoulos, et al., 2012). However, the use of scales in cross-cultural research necessitates certain statistical methods not needed for the analyses of single item measures. Namely, it is, so far, not feasible to statistically test comparability of single-item measures, but there are several statistical methods to test comparability of scales (e.g. Millsap, 2011). According to the TALIS technical reports (OECD, 2014b, 2010), when the

1. The paper presents the results of a study conducted by Jia He as a fellow in the OECD's Thomas J. Alexander Fellowship Programme, in which she was mentored by Katarzyna Kubacka, Analyst in the Directorate for Education and Skills. For more information about the Thomas J. Alexander Fellowship programme, please go to www.oecd.org/edu/thomasjalexanderfellowship.htm

conventional test (i.e. confirmatory factor analysis for comparability) is used, many scales in TALIS do not reach the level of comparability that allows error-free direct comparisons of scale scores across cultures.[2]

Thus, in order to increase the potential of TALIS for policy-relevant cross-cultural research, this paper focuses on the comparability of its scales. It aims to use more flexible tools to test comparability and to locate sources of incomparability, and thereby contributes to more extensive use of the data, maximising the policy implications of these data and informing future planning, administration, and interpretation of international surveys. Specifically, two research objectives are defined, the paper aims to:

1. apply alternative methods to assess the comparability of TALIS scales;

2. systematically assess the level and source of data incomparability.

In the following sections, a short selected review of the literature for each research objective is provided.

2. Literature review

2.1 Bias and invariance

Data incomparability results from bias, which is defined as nuisance factors that jeopardise the validity of instruments applied in different cultures (van de Vijver and Leung, 1997). In cross-cultural research like TALIS, we strive to capture cultural differences while minimising cultural bias. The presence of bias indicates that the scores from the assessment in different cultures reflect some cultural characteristics other than what the assessment is intended to measure. In other words, bias undermines the value of cross-cultural comparisons for policy-making. There are three types of bias depending on the sources of incomparability: (1) *Construct bias*: the construct that is the target of the assessment has a different meaning in different cultures; (2) *Method bias*: there is incomparability due to differences in sampling, respondents' use of the test instruments, and administration modes; and (3) *Item bias*: an item has a different meaning in different cultures.

To check different types of bias, assessment of levels of comparability is needed. For a given scale, the target construct is unobserved, thus it is a latent factor; meanwhile the items of the scale should have certain properties in association with the latent factor to demonstrate comparability. Key parameters include the factor loadings that refer to the association between the item and the latent factor, and item intercepts, which refer to the origin of the item location. Three levels of comparability (also called invariance) in scales can be distinguished:

1. *Configural invariance* means that across cultures the items measuring a construct cover facets of this construct adequately. In statistical terms, this level of invariance signals that items in a measure exhibit the same configuration of the same pattern of zero loadings, loading different from zero, free and fixed parameters in all cultures (Steenkamp and Baumgartner, 1998). As an illustration one can consider a measure of teacher efficacy. If this measure reaches configural invariance, it indicates that teacher efficacy is understood as the same concept and has the same elements across cultures; however, whether the elements relate to the concept with the same strength is not guaranteed.

2. Throughout this paper, the terms "culture" or "cultures" refer to all countries and economies that participated in TALIS.

2. *Metric invariance* indicates that items on the construct have the same factor loadings across cultures. To illustrate, one can consider the measures of temperature using the Fahrenheit and Kelvin scales. The metrics of the two measures are comparable (one unit in Fahrenheit is equal to one unit in Kelvin), whereas the origins of the two measures have a difference of 273 degrees. In the case of teacher efficacy, metric invariance means that each item in this scale is equally related to this construct across cultures. With metric invariance satisfied, scale score comparisons can be made within cultures (e.g. teacher efficacy can be compared between males and females within each culture), and the association of variables can be compared across cultures (e.g. correlations between teacher efficacy and job satisfaction can be compared across cultures, if both teacher efficacy and job satisfaction scales reach metric invariance) (van de Vijver and Tanzer, 2004).

3. *Scalar invariance* implies that items have the same intercepts (i.e. point of origin) across cultures. With the same metric and same origin, scale scores do not have any bias. This would mean in the case of teacher efficacy that all the items are understood and answered in the same way among respondents across all cultures. Thus, scalar invariance would imply that a teacher indicating that he or she is not confident in their ability to teach in Japan, has the same idea of teacher efficacy as one who indicates this in France, for example. Only with scalar invariance can scale scores be validly compared across cultures (van de Vijver and Leung, 1997). This means that sophisticated analyses making use of scale scores across cultures, such as multivariate analyses of variance, structural equation modelling with mean structures, and multilevel analyses, are appropriate only if scalar invariance is established.

The statistical analysis of invariance from the TALIS technical reports revealed that most constructs in TALIS reached metric invariance, but scalar invariance was rarely met (OECD, 2014b, 2010). Previous analyses compared similarities and differences in relationships between constructs in various cultures in terms of metric invariance, (e.g. professional development positively related to teacher efficacy across cultures) and explored within cultural variations (e.g. gender differences in classroom disciplinary climate). So far, scale scores have not been compared directly in TALIS reports because of the lack of full scalar invariance from the conventional test methods. If scalar invariance can be demonstrated, TALIS scales could be used fully for cross-level analyses. As a result, TALIS could be of even more value for cross-cultural research on teaching and learning. Findings from such cross-level and cross-country analyses could serve as compelling evidence for policy making.

2.2 Measurement invariance testing methods

The current practice of invariance tests in TALIS is through multigroup confirmatory factor analysis (OECD, 2014b, 2010), which is a widely used method in cross-cultural research. This analysis uses covariance matrix information to test hierarchical models (i.e. configural, metric, and scalar invariance models). The level of comparability can be inferred from the fit indexes in each model and the comparisons of fit indexes from different models (e.g. Cheung and Rensvold, 2002). The detailed criteria are described in the Annexes where outputs of such analyses are presented.

2.2.1 Criticism of the conventional method

Despite its prevalence in the field of cross-cultural research, recent work has argued that this conventional invariance test method is not well-suited in large-scale international survey contexts. To begin with, the constraints for scalar invariance may be overly strict and unrealistic in comparisons involving dozens of cultures (Lubke and Muthén, 2014). To accept the model of scalar invariance, factor loadings and intercepts of each item of a scale need to be exactly the same across cultures. In the case of TALIS 2013, it would mean factor loadings and intercepts are constrained to be exactly the same across the 38 participating countries and economies; even a slight deviation from one or more cultures would

signal the lack of scalar invariance. Satisfying such demanding constraints is clearly improbable, if not impossible. What is more, from the model fit indexes in conventional tests, it is difficult to tell whether the lack of scalar invariance is caused by major model misspecifications that can lead to erroneous conclusions, or from minor misspecifications that do not have severe consequences for comparability (Byrne and van de Vijver, 2010; Oberski, 2014). Without more detailed information about the cause of incomparability, this psychometric method does not advance our understanding of the source of the incomparability nor facilitate further use of the data. Finally, multigroup confirmatory factor analysis was initially developed for comparisons of two cultures (or three cultures at most). Thus, the criteria for model fit based on a small number of cultures may not apply to comparisons with over two dozens of cultures, as in the case of TALIS (e.g. Rutkowski and Svetina, 2014).

2.2.2 Alternative methods

Due to the limitations in the conventional method, alternative methods aiming at approximate invariance, such as partial invariance and Bayesian estimation, are proposed in the literature. For example, partial invariance means that only a subset of parameters (factor loadings and/or item intercepts) is constrained to be invariant, and the other subset of parameters is allowed to vary across countries. Consequently, the invariant subset can be compared across cultures (Byrne, Shavelson and Muthén, 1989). For instance, if a six-item scale on professional development does not achieve full invariance because one item is understood and answered in different ways in different cultures, then this item can be freely estimated from the scale and the comparisons can be made using the scale with the five equivalent items plus this freely estimated item. Given that most TALIS scales have fewer than five items, and that the comparisons involve over two dozen cultures, the partial invariance approach is not practical.

The more promising approach could be Bayesian approximate invariance testing. Instead of constraining the parameters of loadings and/or intercepts to be exactly the same across cultures, this Bayesian approach allows small differences in these parameters across cultures (Muthén and Asparouhov, 2012; van de Schoot, et al., 2013). The underlying rationale is that although there is no absolute invariance, the slight variations may not severely hinder the comparability, and a valid comparison can still be achieved. Based on previous research (e.g. simulation studies), one can specify a range for loadings and/or intercepts (e.g. loading with a value between 0.49 and 0.51; intercept with a value between 2.49 and 2.51) and check if the data meet the requirement for these ranges. Such a scenario is more realistic in large-scale international surveys, as it provides elasticity/flexibility in the constraints in factor loadings and intercepts. In operational terms, one can specify the pair-wise differences in each parameter (loadings and/or intercepts) across cultures to follow a zero mean and a very small variance (0.01 or 0.05) distribution, in a way to allow some flexibility in these parameters (Lubke and Muthén 2014; Muthén and Asparouhov, 2012). It is suggested that if a model with such specifications fits well, approximate invariance is supported, and the comparisons of scale scores are acceptable (e.g. Cieciuch, et al., 2014; Davidov, et al., 2015). Following recent trends in the literature, Study 1 compares the models of conventional full invariance and Bayesian estimation to investigate whether TALIS scale scores can be validly compared in the absence of a full scalar invariance.

2.3 Sources of the lack of invariance

It should be noted that much effort goes into the development of internationally agreed upon and comparable scales in the preparations for the TALIS main survey. To begin with, the TALIS assessment framework and items are developed through discussions between international groups of experts. Upon the agreement of the experts, rigorous translation, verification, and national adaptions are implemented to balance the comparability and ecological validity of measures. Lastly, these measures are piloted and field trialled, necessary changes are made, and the main study is carried out with standardised administration in paper-and-pencil and computerised survey (OECD, 2014b, 2010). So, the meticulous design and

implementation of TALIS has led to much confidence in the comparability of the scales. In order to further benefit from the wealth of information in TALIS and unlock the potential of cross-cultural scale comparisons, Bayesian approximate invariance testing is proposed. However, if the lack of scalar invariance persists even when using this more flexible method, it is important to understand the root cause and intervene for future rounds of TALIS surveys. Two sources are considered in this paper: (1) the characteristics of the scales under study, and (2) the subgroup of cultures under study.

2.3.1 Scale characteristics and lack of invariance

TALIS enables investigating the link between the lack of scalar invariance and scale characteristics such as length of scale, item length, length of response options, and self-evaluative component in the items. Extant literature demonstrates that scales with simple, unambiguous, and appropriate response anchors can lead to better cross-cultural comparability (Harkness, van de Vijver and Mohler, 2003; van de Vijver and Leung, 1997). In their study on how item complexity and item length are associated with respondents' tendency to always agree with items, Condon, Ferrando, and Demestre (2006) report higher response distortion in items that are long and complex. Thus, the study points to the advantage of shorter and simpler items in getting better comparability.

In addition, Revilla, Saris, and Krosnick (2014) studied the optimal length of response anchors for optimal data quality and comparability. The authors argue that more response options entail enhanced cognitive burden and yield data of lower quality, and they propose to use shorter response scales (e.g. 5-point) rather than longer response scales (e.g. 7- and 11-point). Similarly, Chang (1994) suggests that compared with a 4-point scale, responses on a 6-point scale are more likely to be tainted with measurement errors. This evidence speaks to the benefit of using fewer response options.

Furthermore, topic involvement may also affect how respondents from different cultures respond to surveys (Diamantopoulos, Raeynolds and Simintiras, 2006). For instance, He and van de Vijver (2015) found that the more self-evaluative component in a scale (e.g. evaluating job satisfaction compared with evaluating school climate) leads to higher impression management in survey responding, which may hinder the comparability of data across cultures. Thus, fewer self-evaluative components in a scale can lead to better comparability.

Study 2 investigates TALIS scales with different levels of invariance and the conditions that would stimulate scalar invariance. Based on the above-mentioned literature, Study 2 tests the hypothesis that shorter scales, fewer response options, simpler wording, and fewer self-evaluative components in a scale are associated with better comparability.

2.3.2 Subgroups of cultures and lack of invariance

Besides the scale characteristics, lack of scalar invariance can also stem from differences in the measurements in a few cultures that show particularly pronounced cultural differences in survey responding in contrast to the rest of the countries in a survey. In other words, a measure can be comparable in most TALIS countries and economies, yet the incomparability can stem from large differences in only a subgroup of participating countries and economies (e.g. Harzing, 2006; Johnson, Shavitt and Holbrook, 2011; Yang et al., 2010). For example, it has been argued that East Asian countries and North American countries have very different scale usage preferences. Research shows that East Asians tend to moderate their responses by choosing only the middle categories in a response scale (2 *disagree* and 3 *agree*), whereas North Americans tend to amplify their response by choosing the end points in a response scale (1 *strongly disagree* and 4 *strongly agree*) (Chen, Lee and Stevenson, 1995; Hamamura, Heine and Paulhus, 2008). Subsequently, responding with a 3 (*agree*) on a statement, may actually mean different levels of agreement in different cultures, resulting in lack of scalar invariance across these cultures.

One way to look into the specific cultural variations in measurement is to conduct all pair-wise comparisons. From the pair-wise comparisons, the cultures that show different or similar measurements can be grouped, and conclusions on the general comparability can be made. However, pair-wise comparisons between each and every culture can be extremely cumbersome in the case of TALIS. For instance, in TALIS 2013, pair-wise comparisons of the 38 participating countries and economies would result in a total of 703 comparisons. Another way to pinpoint the specific cultures that show different measurements is a one-to-all comparison (van de Vijver and Leung, 1997). To avoid cumbersomeness, an international average reference group can be extracted from the total sample pool to serve as a reference comparison sample. This can reduce the comparisons to 38 times in the case of TALIS 2013 data.

In Study 3, a pair-wise comparison of measurement invariance of each culture with an international average reference group is carried out to detect country specific variations. It tests whether the lack of overall scalar invariance is caused by only a subset of specific cultures, while the majority of the cultures are comparable.

2.4 Study overview

All in all, three studies are carried out to address the issue of data comparability in the two waves of TALIS. Specifically:

- **Study 1** compares the conventional confirmatory factor analysis method and the Bayesian approximate invariance approach in ascertaining the level of comparability in TALIS principal and teacher self-report scale data. It tests the hypothesis that the Bayesian approximate invariance approach, representing a more realistic picture of a large-scale international survey, shows a higher level of comparability than the conventional method.

- **Study 2** examines the relationship between various scale characteristics and the severity of data incomparability. It is expected that shorter scales, fewer response options, simpler items, and fewer self-evaluative components in a scale are associated with better comparability.

- **Study 3** investigates whether the majority of TALIS participating countries and economies show scalar invariance and the lack of scalar invariance stems from a few countries and economies that have a sharper difference in survey responding behaviours. The scale on *Satisfaction with Current Work Environment* is used as an illustration of the new proposed method in this study.

3. Study 1: Comparisons of full and approximate invariance

To check the comparability of scales of principal and teacher self-reports in the two rounds of TALIS surveys (i.e. TALIS 2008 and TALIS 2013), the model fits of scales from the conventional full measurement invariance testing and the Bayesian approximate invariance are compared. It is expected that the latter provides a more flexible outlook on measurement invariance issues and leads to a better fit.

3.1 Method

Data source

In TALIS 2008, 24 countries and economies participated in the survey, and data on 22 countries and economies are available for the analysis.[3] In TALIS 2013, initially 34 countries and economies participated

3. Iceland withdrew the data to protect respondents' privacy and the Netherlands had an overall participating rate of 16.7%, which was significantly below the international requirement for response rates, thus these two countries were not included in the analysis.

and 4 additional countries and economies (i.e. China, Shanghai; Georgia; New Zealand; and the Russian Federation) collected data one year later than the main field study, thus data from 38 countries and economies are included in the analysis.[4] The core survey targets lower secondary level principals and teachers (ISCED 2), and this project focuses on ISCED 2 principals and teachers. Given the variations in sample sizes in each participating country or economy, which may have an effect on the comparability testing, a subsample of 1 000 teachers and 150 principals are randomly selected from each culture for the 2008 TALIS (round 1), and a subsample of 1 500 teachers and 150 principals selected for the 2013 TALIS participating countries and economies (round 2). The subsample size of teachers was set as 1 500 because of the larger sample size in all the participating countries and economies in 2013. As the technical reports suggest (OECD, 2014b, 2010), this selected sample is large enough to yield meaningful comparisons.

Analyses. As already documented in the technical reports (OECD, 2014b, 2010), scalar invariance through conventional multigroup confirmatory factor analysis is not achieved in the comparisons of all available participating countries and economies in TALIS 2008, nor for the 34 countries and economies in 2013. In this study, the comparability analysis as reported in the technical reports is replicated and extended in two ways. First, data from four additional countries and economies (i.e. China, Shanghai; Georgia; New Zealand; and the Russian Federation) are included in the second round of TALIS; data from these countries and economies were not available for analysis when the 2013 TALIS technical report was produced. Therefore, the comparability of scales across 38 countries and economies is checked. Second, the responses to the items in each scale are treated first as continuous, then as ordered categories (Desa, 2014).[5] Thus, two analyses are carried out for each scale. Sample syntaxes for the two analyses are provided in Annex 1.

For the Bayesian approximate invariance testing, the pair-wise difference in all loadings and intercepts is set as following a distribution of zero mean and 0.05 variance (sample syntax for the estimation is provided in Annex 1). All the analyses are carried out with Mplus 7.0 (Muthén and Muthén, 1998-2012).

3.2 Results

Consistent with the technical reports (OECD, 2014b, 2010), with conventional confirmatory factor analysis as a comparability test, none of the scales in principal or teacher self-reports show full scalar invariance[6]. As the results are similar to the TALIS technical reports, only the summary of fit indexes in the 2013 TALIS teacher self-report scales across all 38 countries and economies is presented in Annex 2.

Next, the results of applying the Bayesian approximate invariance testing to each principal and teacher self-report scale, as a more flexible method, are presented in Table 1. A summary of model fit for all scales is presented in Annex 3. In general, three scales in the TALIS 2008 principal self-report – *Constructivist Beliefs about Instruction*, *Accountability Role of the Principal*, and *Promoting Instructional Improvements and Professional Development* – show acceptable approximate invariance. Four principal scales in TALIS 2013 – *Distributed Leadership*, *Satisfaction with Profession*, *Instructional Leadership*,

4. In TALIS 2013, the United States did not reach the international requirement for response rates, but it had an overall participation rate of 51.4%, so it was included in the analysis.

5. The treatment of responses on items as "ordered categories" in comparison to treatment as "continuous" is an improvement in comparability testing and is expected to improve the model fit within the multigroup confirmatory factor analysis framework.

6. In some cases, treating responses as "ordered categorical" helps to improve the model fit, mainly in the metric invariance level, but the full scalar invariance is still not reached. In some other cases, the model does not converge when the data are treated as ordered categorical, thus it does not improve the comparability testing results.

and *Mutual Respect* – show acceptable approximate invariance. In contrast, regarding teacher self-reports, only one scale, namely, *Classroom Disciplinary Climate*, in TALIS 2008, reaches approximate invariance.

Therefore, as hypothesised, the Bayesian approximate invariance approach yields better results than the conventional multigroup confirmatory factor analysis in establishing measurement invariance of scales in TALIS. This is especially true for principal self-reports. At the same time, results show a persistent lack of approximate invariance in intercepts in the teacher scales. This means that within-cultural comparisons of scale means and cross-cultural comparisons of correlations or regressions are feasible and reasonable, whereas direct cross-cultural comparisons of scale means may suffer from many measurement errors. This lack of invariance may be a result of more variations in teachers' backgrounds across cultures, compared with principals (OECD, 2014b).

Table 1. Scales that show approximate invariance

2008 Principal Self-Reports
Constructivist Beliefs about Instruction
Accountability Role of the Principal
Promoting Instructional Improvements and Professional Development
2013 Principal Self-Reports
Distributed Leadership
Satisfaction with Profession
Instructional Leadership
Mutual Respect
2008 Teacher Self-Reports
Classroom Disciplinary Climate

4. Study 2: Linking scale characteristics with lack of invariance

Study 2 investigates the association between scale characteristics and scale comparability. The purpose of this study is to reveal which scales characteristics are more likely to be associated with scale invariance.

4.1 Method

Data source

Scales of ISCED 2 principal and teacher self-reports in the two rounds of TALIS surveys are used in this study. The items forming each scale and the corresponding response anchors can be found in the TALIS technical reports (OECD, 2014b, 2010).

Analysis

The conventional full measurement invariance testing for each scale is carried out in multigroup confirmatory factor analysis in Mplus 7.0 (Muthén and Muthén, 1998-2012), as in Study 1. The threat to incomparability in each scale is operationalised as the change of the comparative fit index (CFI) value from the metric to the scalar invariance model. The CFI is an index with a value from 0 to 1, and a higher value indicates a better model fit. CFI values generally decrease from a more liberal to a more constrained model (i.e. from the metric to scalar invariance model) (Cheung and Rensvold, 2002). As most scales in

TALIS reach metric invariance, the CFI value in the metric invariance model was used as the baseline, and the drop in the CFI value from the metric to scalar invariance model can serve as a proxy of how severe the lack of scalar invariance is.

The item content and the response options of each scale were documented and features were recorded for the analysis. The analyses focus on four scale characteristics:

1. Number of items in one scale.

2. Item length, operationalised as the average number of words in each item in a given scale. In the present study, the English version of items are used.

3. Response option length (Likert-scale points).

4. Self-evaluative component, coded as 1 when all items in the scale are about self-evaluation (e.g. self-efficacy), 2 when items are a mix of self and other evaluation (e.g. student-teacher relationship), and 3 when items are all about other or contextual evaluation (e.g. school climate).

4.2 Results

Table 2 below presents an overview of the scale characteristics and level of lack of scalar invariance for principal and teacher self-report scales in 2008 and in 2013, respectively, in the descending order of comparability (i.e. a change in the CFI indicates the CFI value change from the metric invariance to the scalar invariance model, which is the indicator of severity of data incomparability).

Table 2. Overview of scale characteristics and lack of invariance indicator

Scales	No. of Items	Item length	Response length	Self-evaluation	Change CFI
2008 Principal Self-Reports					
School Climate: Student Delinquency	6	6	4	3	0.14
Promoting Instructional Improvements and Professional Development	4	12	4	1	0.28
Framing and Communicating the School's Goals and Curricular Development	6	13	4	1	0.32
School Climate: Teachers' Working Morale	3	3	4	3	0.32
Constructivist Beliefs about Instruction	4	14	4	2	0.48
Bureaucratic Rule-Following	5	15	4	1	0.51
Accountability Role of the principal	4	21	4	1	0.68
Supervision of the Instruction in the School	4	9	4	1	0.93
2013 Principal Self-Reports					
Satisfaction with Current Work Environment	4	9	4	1	0.14
Satisfaction with Profession	4	10	4	2	0.18
Instructional Leadership	3	14	4	1	0.19
School Delinquency and Violence	4	8	5	3	0.20
Distributed Leadership	3	13	4	3	0.25
Mutual Respect	4	8	4	3	0.28

Scales	No. of Items	Item length	Response length	Self-evaluation	Change CFI
2008 Teacher Self-Reports					
Classroom Disciplinary Climate	4	12	4	2	0.05
Teacher-Student Relations	4	13	4	3	0.12
Self-Efficacy	4	13	4	1	0.15
Classroom Teaching Practice: Enhanced Activities	4	16	6	2	0.27
Classroom Teaching Practice: Student-Oriented	4	14	6	2	0.30
Constructivist Beliefs about Instruction	4	14	4	2	0.47
Classroom Teaching Practice: Structuring	5	10	6	2	0.57
Professional Collaboration	5	9	6	1	0.79
Exchange and Co-ordination for Teaching	5	9	6	1	0.83
Direct Transmission Beliefs about Instruction	4	15	4	3	0.93
2013 Teacher Self-Reports					
Classroom Disciplinary Climate	4	7	4	2	0.03
Efficacy in classroom management	4	7	4	1	0.10
Satisfaction with Current Work Environment	4	10	4	1	0.10
Participation among Stakeholders	5	12	4	3	0.11
Needs for Professional Development in Subject Matter and Pedagogy	5	6	4	1	0.11
Efficacy in Student Engagement	4	7	4	1	0.13
Teacher-Student Relations	4	6	4	3	0.15
Efficacy in Instruction	4	7	4	1	0.17
Satisfaction with Profession	4	6	4	2	0.18
Need for Professional Development for Teaching for Diversity	6	7	4	1	0.20
Constructivist Beliefs	4	7	4	2	0.26
Effective Professional Development	4	6	4	3	0.75
Exchange and Coordination for Teaching	4	9	6	1	0.76
Professional Collaboration	4	7	6	1	0.99

Note: The change in CFI must be less than 0.01 for the scalar invariance model to be preferred.

To further study the associations between the scale characteristics and lack of invariance, Spearman's rank correlations are produced for each set of principal and teacher self-reports in 2008 and 2013 (Table 3), respectively. Spearman’s rank correlation is used, because the CFI values are not normally distributed.

Table 3. Relationship between lack of scalar invariance and scale characteristics (Spearman's rank correlation)

Lack of scalar invariance	No. of items	Item length	Response length	Self-evaluation
2008 Principal Data (n =8)	-0.38	0.45	Na	-0.45
2013 Principal Data (n =6)	-0.21	-0.26	0.13	0.83*
2008 Teacher Data (n =10)	0.57	-0.15	0.38	-0.14
2013 Teacher Data (n =14)	-0.12	-0.18	0.61*	-0.02

Note: *$p < .05$

Given the limited number of scales and the diversity in scale compositions, interpretation of these correlations should be cautious. The most consistent predictor across the principal and teacher self-reports of lack of scalar invariance is response length: more response options (e.g. 6-point Likert scale) tend to hinder scalar comparability. This finding is in accordance with previous studies (Chang, 1994; Revilla, Saris and Krosnick, 2014). The other three scale characteristics show mixed results. For principal scales, scales with fewer items actually show worse comparability compared with scales with more items. For teacher scales, scales of self-evaluation seem to lead to worse comparability, whereas item length does not seem to lead to lower level of scalar invariance.

5. Study 3: Linking specific cultures with lack of invariance

Study 3 investigates specific cultural variations through a pair-wise comparison of measurement invariance between each culture and an international average reference group. This international average reference group can be a pool of a subsample from each culture, and this pooled sample represents what the teacher force looks like across all participating countries and economies. This reference group is meant to avoid the cumbersome pair-wise comparisons of the 38 participating countries and economies (which would result in a total of 703 comparisons). The scale on *Satisfaction with Current Work Environment* is used as an illustration. This study aims to uncover the cultural clusters that can be compared validly and pinpoint cultures that show measurement differences, and in general assess the scope of data incomparability in scales. In addition, the Study presents a novel method of cultural comparison using the average reference group in TALIS.

5.1 Method

Data source. The ISCED 2 teacher self-report data of the *Satisfaction with Current Work Environment* scale from all the 38 participating countries and economies are used. This scale has four items ("I would like to change to another school if that were possible"; "I enjoy working at this school"; "I would recommend my school as a good place to work", and "All in all, I am satisfied with my job"). The response options range from (1) *strongly disagree* to (4) *strongly agree*.

Analysis. An international average reference group was extracted from the total sample pool to serve as a comparison sample. For this purpose, a random sample of 200 teachers was selected from each of the TALIS countries or economies and pooled together to represent an international average reference group. Thus this reference group has a total sample of 7 600 teachers coming from 38 cultures. Next, a conventional multigroup confirmatory factor analysis was performed for each culture (excluding the 200 teachers who were assigned to the reference group) and the international average reference group. So, the comparisons were made 38 times.

5.2 Results

For 27 out of the 38 cultures, the comparisons with the international average reference group show full scalar invariance. This means that these 27 cultures are comparable to the international average, and thus they should be comparable to each other on the *Satisfaction with Current Work Environment* scale. The cultures that show a lack of scalar invariance with the international average reference group, and thus should not be used in cross-cultural comparisons using TALIS, are listed in Table 4. These cultures clearly belong to three cultural clusters: East Asian, European, and Latin American. It implies that these cultural clusters show very different survey responding behaviours, thus making themselves different to the other cultures in the sample. From previous studies on response styles in different cultures, it could be that the East Asian cluster has a higher tendency to show a modesty bias (i.e. always endorsing the midpoints of a scale), whereas the European and the Latin American cluster seem to have a higher tendency to show self-enhancement (i.e. always endorsing the end points of a scale). The summary of the model fit indexes for all the 38 comparisons is presented in Annex 4. It is interesting to carry out the same analyses with other scales and check whether the same clusters of cultures always stand out in the comparability test. This study serves as an example for such analyses.

Table 4. Cultures that show lack of scalar invariance when comparing with the international average reference group

Cultural Clusters	Cultures
East Asia	Malaysia, Singapore
Europe	Czech Republic, Croatia, Estonia, Latvia, Slovak Republic, England (United Kingdom)
Latin America	Brazil, Chile, Mexico

6. Conclusions and implications

A demonstration of data comparability is a prerequisite for any cross-cultural comparison. TALIS has been the first large-scale international survey that formally tested data comparability (OECD 2014, 2010). Given the large number of countries and economies involved in the survey, and the overly strict conventional statistical method to test comparability, many scales in TALIS have not reached the level of comparability that allows direct comparisons of scale scores. To facilitate effective data analysis of TALIS and maximise its policy implications, this report proposes to (1) use a more flexible statistical method to test comparability, and to (2) understand the level and sources of scale data incomparability in three studies. A summary of the findings and the implications for each study is provided in this section.

6.1 A more flexible statistical method for comparability testing

Study 1 compared the conventional statistical method with the more flexible Bayesian approximate invariance test in data comparability testing. The results show that while none of the scales achieves full comparability using the conventional method, over half of the principal scales in both the first and second round of TALIS reach approximate invariance using the Bayesian approach. Thus these scale scores can be validly compared across cultures. However, most teacher scales do not reach approximate invariance, and great caution is needed when these scale scores are compared across cultures.

Although not all scales reach approximate invariance using the Bayesian testing, the present findings show that this approach is superior to the conventional comparability testing method in large-scale international survey contexts. The Bayesian approach takes into consideration the possibility that there is a large amount of very small measurement variations across cultures, and allows the small, trivial variations in the testing (Davidov, et al., 2015; Lubke and Muthén, 2014; Muthén and Asparouhov 2012). Therefore it is a more flexible tool for large-scale surveys. However, it also should be noted that this approach is

relatively new and more empirical and simulation studies are needed to affirm the prior specifications appropriate for large-scale survey data. Thus, it is recommended that it be carried out side by side with the conventional testing for future rounds of TALIS data comparability testing and for other large-scale international surveys.

6.2 Fewer response options and enhanced comparability

Study 2 investigated the relationship between data comparability and four scale characteristics (e.g. scale length, item length, number of response options, and self-evaluative components) in TALIS 2008 and 2013 principal and teacher scales. The results show that scales with shorter response options (e.g. a 4-point scale instead of a 6-point scale) generally have better comparability. The other three scale characteristics do not show consistent patterns. Therefore, a clear message for future TALIS, as well as similar large-scale surveys, is to keep the response options short, in order to improve data comparability. At the same time, the survey development process will need to balance the need for data comparability with that of investigating more nuanced phenomena and the compromise between the two is likely to ultimately determine the scale length.

6.3 Comparability within the majority of cultures and lack of comparability in a few cultures

Study 3 analysed the specific cultural variations that can account for lack of data comparability in TALIS surveys. Namely, it tested the comparability of the *Satisfaction with Current Work Environment* scale between each participating country or economy with a pooled international average reference group. The findings show that all the TALIS cultures reach metric invariance, which indicates the relationship between variables can be validly compared across cultures. Moreover, the results show that 27 out of 38 countries and economies demonstrate full comparability with the international average reference group. Subsequently, cross-cultural comparisons among these countries and economies using scale scores of the *Satisfaction with Current Work Environment* scale are justified. The countries and economies that do not show full comparability with the international average reference group (i.e. Brazil, Chile, Croatia, the Czech Republic, England-United Kingdom, Estonia, Latvia, Malaysia, Mexico, Singapore and the Slovak Republic) may differ significantly in their languages and survey responding behaviours. Therefore, when these countries and economies are involved in scale score comparisons, caution in interpretation is much needed.

Moreover, Study 3 used a single TALIS scale (*Satisfaction with Current Work Environment)* as an illustration of a possible new method to establish comparability. The results provide evidence that pair-wise comparisons of cultures with a pooled international average reference group can be applied to other constructs. This method could serve as a more practical method of establishing comparability for TALIS and other large-scale surveys and a way to pinpoint the cultures that are more different in survey responding than others.

6.4 Conclusions

To summarise, further cross-cultural research on TALIS data to maximise its policy potential is contingent on solid research methods and valid comparisons. This project contributes to better understanding the level of scale data comparability in the two rounds of TALIS in a number of ways. Results from the three studies show that more flexible tools in comparability testing are necessary in order to meet the new challenges in large-scale surveys. The Bayesian approximate invariance test used in this project seems a good candidate for demonstrating scale score comparability, while other, newer approaches, such as alignment (Asparouhov and Muthén, 2014) and multilevel models (Jak, Oort and Dolan, 2013; Jak, Oort and Dolan, 2014), may offer additional insights into the comparability issue. These

novel methods should be experimented within TALIS and other large-scale international surveys, in order to find the most suitable methods for data comparability in surveys with the large scope that TALIS has.

Regarding the lack of comparability, this paper demonstrates that scales with fewer response options generally show better comparability (Study 2). Thus, future TALIS and other survey design should take this into consideration when constructing scale response anchors. The present work also finds that the majority of TALIS cultures are comparable, whereas incomparability stems from a few cultures in the *Satisfaction with Current Work Environment* scale (Study 3). This suggests the potential to carry out valid scale score comparisons in most cultures and the caution needed in a few cultures with large variations in survey responding behaviours. The comparison between each culture and a pooled international average reference group is a relatively easy way to detect the scope of incomparability from specific cultures, in surveys with as many countries as TALIS has. All in all, the findings of this project could be generalised and applied not only to the future TALIS cycles, but also to other large-scale surveys, in order to boost the potential of large scale data for evidence-based policy making.

REFERENCES

Asparouhov, T. and B. Muthén (2014), "Multiple-group factor analysis alignment", *Structural Equation Modeling: A Multidisciplinary Journal*,Vol. 21/3, pp. 495-508.

Byrne, B.M., R.J. 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, pp. 456-466.

Byrne, B.M. and F.J.R. van de Vijver (2010), "Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence", *International Journal of Testing*, Vol. 10/2, pp. 107-132.

Chang, L. (1994), "A Psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity", *Applied Psychological Measurement*, Vol. 18/3, pp. 205-215.

Chen, C., S.-Y Lee and H. W. Stevenson (1995), "Response style and cross-cultural comparisons of rating scales among East Asian and North American students", *Psychological Science*, Vol. 6/3, pp. 170-175.

Chen, F. F. (2008), "What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research", *Journal of Personality and Social Psychology*, Vol. 95/5, pp. 1005-1018.

Cheung, G. W. and R. B. Rensvold (2002), "Evaluating goodness-of-fit indexes for testing measurement invariance", *Structural Equation Modeling*, Vol. 9/2, pp. 233-255.

Cieciuch, J., et al. (2014), "Comparing results of an exact vs. an approximate (Bayesian) measurement invariance test: a cross-country illustration with a scale to measure 19 human values", *Frontiers in Psychology*, Vol. 5, p. 982.

Condon, L., P. J. Ferrando and J. Demestre (2006), "A note on some item characteristics related to acquiescent responding", *Personality and Individual Differences,*Vol. 40/3, pp. 403-407.

Davidov, E., et al. (2015), "The comparability of measurements of attitudes toward immigration in the European Social Survey: Exact versus approximate measurement equivalenc", *Public Opinion Quarterly* , Vol. 79/S1, pp. 244-266.

Desa, D. (2014), "Evaluating measurement invariance of TALIS 2013 complex scales: Comparison between continuous and categorical multiple-group confirmatory factor analyses", *OECD Education Working Papers*, No.103, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jz2kbbvlb7k-en.

Diamantopoulos, A., et al. (2012), "Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective", *Journal of the Academy of Marketing Science*, Vol. 40/3, pp. 434-449.

Diamantopoulos, A., N. L. Raeynolds and A. C. Simintiras (2006), "The impact of response styles on the stability of cross-national comparisons", *Journal of Business Research,* Vol. 59/8, pp. 925-935.

Hamamura, T., S. J. Heine and D. L. Paulhus (2008), "Cultural differences in response styles: The role of dialectical thinking", *Personality and Individual Differences*, Vol. 44/4, pp. 932-942.

Harkness, J. A., F. J. R. van de Vijver and P. P. Mohler (eds.) (2003), *Cross-Cultural Survey Methods*, John Wiley & Sons, Hoboken, New Jersey.

Harzing, A.-W. (2006), "Response styles in cross-national survey research: A 26-country study", *International Journal of Cross Cultural Management*, Vol. 6/2, pp. 243-266.

He, J. and F. J. R. van de Vijver (2015), "Effects of a general response style on cross-cultural comparisons: Evidence from the Teaching and Learning International Survey", *Public Opinion Quarterly*, Vol. 79/S1, pp. 267-290.

Jak, S., F. J. Oort and C. V. Dolan (2014), "Measurement bias in multilevel data", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 21/1, pp. 31-39.

Jak, S., F. J. Oort and C. V. Dolan (2013), "A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 20/2, pp. 265-282.

Johnson, T. P., S. Shavitt and A. L. Holbrook (2011), "Survey response styles across cultures", in D. Matsumoto and F. J. R; van de Vijver (eds.), *Cross-Cultural Research Methods in Psychology*, Cambridge University Press, New York, NY, pp. 130-175.

Lubke, G. H. and B. O. Muthén (2014), "Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 11/4, pp. 514-534.

Meredith, W. (1993), "Measurement invariance, factor analysis and factorial invariance", *Psychometrika*, Vol. 58/4, pp. 525-543.

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

Muthén, B. and T. Asparouhov (2012), "Bayesian SEM: A more flexible representation of substantive theory", *Psychological Methods*, Vol. 17/3, pp. 313-335.

Muthén, L. K. and B. O. Muthén (1998-2012), *Mplus User's Guide, Seventh Edition*, Muthén & Muthén, Los Angeles, CA.

Oberski, D. L. (2014), "Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models", *Political Analysis*, Vol. 22/1, pp. 45-60.

OECD (2014a), *TALIS 2013 Results: An International Perspective on Teaching and Learning*, TALIS, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264196261-en.

OECD (2014b), *TALIS 2013 Technical Report*, OECD Publishing, Paris, www.oecd.org/edu/school/TALIS-technical-report-2013.pdf.

OECD (2010), *TALIS 2008 Technical Report*, TALIS, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264079861-en.

Revilla, M. A., W. E. Saris and J. A. Krosnick (2014), "Choosing the number of categories in agree–disagree scales", *Sociological Methods & Research* , Vol. 43/1, pp. 73-97.

Rutkowski, L. and D. Svetina (2014), "Assessing the hypothesis of measurement invariance in the context of large-scale international surveys", *Educational and Psychological Measurement* , Vol. 74/1, pp. 31-57.

Steenkamp, J.-B. E. M. and H. Baumgartner (1998), "Assessing measurement invariance in cross-national consumer research", *Journal of Consumer Research* , Vol. 25/1, pp. 78-107.

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.

van de Vijver, F. J. R. and K. Leung (1997), *Methods and Data Analysis of Comparative Research*, Sage, Thousand Oaks, CA.

van de Vijver, F. J. R. and N. K. Tanzer (2004), "Bias and equivalence in cross-cultural assessment: An overview", *European Review of Applied Psychology*, Vol. 54/2, pp.119-135.

Yang, Y., et al. (2010), "Response styles and culture" in J.A. Harkness, et. al. (eds.), *Survey Methods in Multinational, Multiregional and Multicultural Contexts*, Wiley, New York, NY, pp. 203-223.

ANNEX 1: SAMPLE SYNTAX

Figure 1. Sample syntax for the *Satisfaction with Current Work Environment* scale for the conventional invariance testing in Mplus (treating data as continuous) in Study 1[7]

```
DATA:
 FILE IS "Satisfaction with Current Work Environment.dat";

VARIABLE:
  NAMES ARE CountryID  TT2G46C TT2G46E TT2G46G TT2G46J;
  USEVARIABLES = TT2G46C TT2G46E TT2G46G TT2G46J;
  MISSING = TT2G46C(7, 8, 9);
  MISSING = TT2G46E(7, 8, 9);
  MISSING = TT2G46G(7, 8, 9);
  MISSING = TT2G46J(7, 8, 9);
  GROUPING = CountryID(1=Austrailia 2=Brazil 3=Bulgaria 4=Chile 5=Croatia 6=Cyprus  7=Czech
8=Denmark 9=Estonia 10=Finland 11=France 12=Iceland 13=Israel 14=Italy 15=Japan 16=Korea
17=Latvia 18=Malaysia 19=Mexico 20=Netherlands  21=Norway 22=Poland 23=Portugal 24=Romania
25=Serbia 26=Singapore 27=Slovak 28=Spain 29=Sweden 30=AbuDhabi 31=Alberta 32=England
33=Belgium 34=United States 35=Georgia 36=NewZealand 37=Russia 38=Shanghai);

ANALYSIS:
ESTIMATOR = MLR;
MODEL = CONFIGURAL METRIC SCALAR
MODEL:
f1 by TT2G46C TT2G46E TT2G46G TT2G46J;

OUTPUT:  SAMPSTAT STDYX ;
```

7. 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 shall preserve 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.

Figure 2. Sample syntax for the *Satisfaction with Current Work Environment* scale for the conventional invariance testing in Mplus (treating data as categorical) in Study 1[8]

```
DATA:
  FILE IS "Satisfaction with Current Work Environment.dat";

VARIABLE:
  NAMES ARE SchoolID CountryID  TT2G46C TT2G46E TT2G46G TT2G46J
 MODE TCHWGT;
CATEGORICAL= TT2G46C TT2G46E TT2G46G TT2G46J;
  USEVARIABLES = TT2G46C TT2G46E TT2G46G TT2G46J;
  MISSING = TT2G46C(7, 8, 9);
  MISSING = TT2G46E(7, 8, 9);
  MISSING = TT2G46G(7, 8, 9);
  MISSING = TT2G46J(7, 8, 9);

   GROUPING = CountryID(1=Austrailia 2=Brazil 3=Bulgaria 4=Chile 5=Croatia 6=Cyprus  7=Czech
8=Denmark 9=Estonia 10=Finland 11=France 12=Iceland 13=Israel 14=Italy 15=Japan 16=Korea
17=Latvia 18=Malaysia 19=Mexico 20=Netherlands  21=Norway 22=Poland 23=Portugal 24=Romania
25=Serbia 26=Singapore 27=Slovak 28=Spain 29=Sweden 30=AbuDhabi 31=Alberta 32=England
33=Belgium 34=United States 35=Georgia 36=NewZealand 37=Russia 38=Shanghai);

ANALYSIS:
ESTIMATOR = WLSM;
MODEL = CONFIGURAL METRIC SCALAR
MODEL:
f1 by TT2G46C TT2G46E TT2G46G TT2G46J

OUTPUT:  SAMPSTAT STDYX;
```

8. See note 7.

Figure 3. Sample syntax for the *Satisfaction with Current Work Environment* scale for the Bayesian approximate invariance testing in Mplus in Study 1

```
DATA:
FILE IS "Satisfaction with Current Work Environment.dat";

VARIABLE:
 NAMES ARE CountryID TT2G46C TT2G46E TT2G46G TT2G46J;
  USEVARIABLES = TT2G46C TT2G46E TT2G46G TT2G46J;
  MISSING = TT2G46C(7, 8, 9);
  MISSING = TT2G46E(7, 8, 9);
  MISSING = TT2G46G(7, 8, 9);
  MISSING = TT2G46J(7, 8, 9);
  KNOWNCLASS is G(CountryID =1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
   21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38);
  CLASSES is G(38);

ANALYSIS:
    MODEL = ALLFREE;
    TYPE = MIXTURE;
    ESTIMATOR = BAYES;
    BSEED = 200;
    CHAINS = 4;
    BITERATIONS = 50000(10000);
    BCONVERGENCE = 0.01;
     PROCESSORS = 4;

MODEL:
%overall%
f1 by TT2G46C* TT2G46E TT2G46G TT2G46J (lam#_1-lam#_4);
f1@1;
[f1@0];
[TT2G46C TT2G46E TT2G46G TT2G46J] (nu#_1 - nu#_4);

MODEL PRIORS:
do(1, 4) diff(lam1_# - lam38_#) ~ N(0, .05);
do(1, 4) diff(nu1_# - nu38_#) ~ N(0, .05);

OUTPUT: tech1 tech8;

PLOT: type is plot2;
```

ANNEX 2: SUMMARY OF FIT INDEXES IN 2013 TALIS TEACHER SELF-REPORT SCALES IN THE CONVENTIONAL MULTIGROUP CONFIRMATORY FACTOR ANALYSIS FRAMEWORK IN STUDY 1

Classroom Disciplinary Climate			CFI	TLI	RMSEA
	Continuous	Configural	1.00	0.99	0.08
		Metric	0.97	0.97	0.13
		Scalar	0.94	0.95	0.15
	Categorical		Too few categories, thus not possible		
Constructivist Belief			CFI	TLI	RMSEA
	Continuous	Configural	0.98	0.95	0.07
		Metric	0.97	0.96	0.06
		Scalar	0.71	0.78	0.14
	Categorical	Configural	0.99	0.98	0.09
		Metric	0.99	0.99	0.07
		Scalar	0.93	0.96	0.13
Effective Professional Development			CFI	TLI	RMSEA
	Continuous	Configural	0.98	0.95	0.07
		Metric	0.96	0.95	0.07
		Scalar	0.78	0.83	0.12
	Categorical	Configural	0.99	0.98	0.10
		Metric	0.96	0.95	0.15
		Scalar	0.92	0.96	0.13
Efficacy in Classroom Management			CFI	TLI	RMSEA
	Continuous	Configural	0.99	0.96	0.08
		Metric	0.98	0.98	0.06
		Scalar	0.88	0.91	0.12
	Categorical	Configural	no convergence		
		Metric	1.00	0.99	0.08
		Scalar	0.98	0.99	0.13
Efficacy in Instruction			CFI	TLI	RMSEA
	Continuous	Configural	0.99	0.97	0.06
		Metric	0.98	0.97	0.06
		Scalar	0.81	0.85	0.14
	Categorical	Configural	1.00	0.99	0.07

		Metric	0.98	0.98	0.11
		Scalar	0.96	0.97	0.14
Efficacy in Student Engagement			CFI	TLI	RMSEA
	Continuous	Configural	0.98	0.93	0.10
		Metric	0.97	0.96	0.07
		Scalar	0.84	0.88	0.13
	Categorical	Configural	1.00	0.98	0.12
		Metric	0.99	0.98	0.12
		Scalar	0.97	0.98	0.13
Exchange and Coordination for Teaching			CFI	TLI	RMSEA
	Continuous	Configural	0.99	0.96	0.07
		Metric	0.95	0.93	0.09
		Scalar	0.19	0.38	0.26
	Categorical	Configural	1.00	0.99	0.08
		Metric	no convergence		
		Scalar	0.55	0.83	0.27
Need for Professional Development for Teaching for Diversity			CFI	TLI	RMSEA
	Continuous	Configural	0.89	0.82	0.14
		Metric	0.87	0.86	0.12
		Scalar	0.67	0.74	0.16
	Categorical	Configural	0.97	0.94	0.19
		Metric	no convergence		
		Scalar	0.91	0.94	0.19
Needs for Professional Development in Subject Matter and Pedagogy			CFI	TLI	RMSEA
	Continuous	Configural	0.94	0.88	0.13
		Metric	0.91	0.90	0.12
		Scalar	0.80	0.84	0.15
	Categorical	Configural	0.99	0.98	0.18
		Metric	no convergence		
		Scalar	0.97	0.98	0.16
Participation among Stakeholders			CFI	TLI	RMSEA
	Continuous	Configural	0.90	0.79	0.16
		Metric	0.88	0.87	0.13
		Scalar	0.78	0.83	0.15
	Categorical	Configural	0.86	0.73	0.70
		Metric	no convergence		
		Scalar	0.97	0.98	0.17

Professional Collaboration			CFI	TLI	RMSEA
	Continuous	Configural	0.99	0.96	0.05
		Metric	0.93	0.91	0.07
		Scalar	0.00	0.07	0.24
	Categorical	Configural	0.99	0.98	0.06
		Metric	0.75	0.70	0.23
		Scalar	0.00	0.61	0.26
Satisfaction with Current Work Environment			CFI	TLI	RMSEA
	Continuous	Configural	0.99	0.96	0.07
		Metric	0.97	0.96	0.07
		Scalar	0.87	0.90	0.11
	Categorical	Configural	1.00	0.99	0.11
		Metric	1.00	0.99	0.09
		Scalar	0.98	0.99	0.11
Satisfaction with Profession			CFI	TLI	RMSEA
	Continuous	Configural	0.95	0.86	0.13
		Metric	0.94	0.92	0.10
		Scalar	0.75	0.81	0.15
	Categorical	Configural	0.99	0.97	0.21
		Metric	0.99	0.98	0.16
		Scalar	0.97	0.98	0.15
Teacher-Student Relations			CFI	TLI	RMSEA
	Continuous	Configural	0.99	0.98	0.05
		Metric	0.98	0.98	0.05
		Scalar	0.83	0.87	0.13
	Categorical	Configural	1.00	1.00	0.08
		Metric	0.99	0.99	0.09
		Scalar	0.98	0.98	0.14

Notes:

1. The model fit was evaluated by the Tucker Lewis Index (TLI) (acceptable above 0.90), Comparative Fit Index (CFI) (acceptable above 0.90), and Root Mean Square Error of Approximation (RMSEA) (acceptable below 0.08).

2. The acceptance of a more constrained model is based on the change of CFI (acceptable within 0.01 from a less to a more constrained model) (Cheung and Rensvold, 2002; Rutkowski and Svetina, 2014).

ANNEX 3: SUMMARY OF MODEL FIT IN THE BAYESIAN APPROXIMATE INVARIANCE TESTS IN STUDY 1

Scales	PPP	CI lower	CI higher
2008 Principal Self-Reports			
Constructivist Beliefs about Instruction	0.49	-69.44	68.17
Accountability Role of the Principal	0.31	-53.14	80.33
Promoting Instructional Improvements and Professional Development	0.22	-39.10	95.17
School Climate: Teachers' Working Morale	0.01	14.93	141.22
Supervision of the Instruction in the School	0	11.40	152.14
Bureaucratic Rule-Following scale	0	120.72	284.54
Framing and Communicating the School's Goals and Curricular Development	0	399.58	578.83
School Climate: Student Delinquency	0	481.38	664.65
2013 Principal Self-Reports			
Distributed Leadership	0.45	-70.80	78.65
Satisfaction with Profession	0.17	-39.85	113.50
Instructional Leadership	0.16	-37.19	110.66
Mutual Respect	0.06	-13.41	163.55
School Delinquency and Violence	0	74.65	263.89
Satisfaction with Current Work Environment	0	975.54	1155.31
2008 Teacher Self-Reports			
Classroom Disciplinary Climate	0.04	-8.11	125.98
Constructivist Beliefs about Instruction	0	23.66	157.40
Direct Transmission Beliefs	0	61.78	199.59
Classroom Teaching Practice: Structuring	0	69.12	230.75
Teacher-Student Relations	0	115.33	252.84
Classroom Teaching Practice: Student-Oriented	0	213.45	354.08
Self-Efficacy	0	442.44	580.08
Exchange and Co-ordination for Teaching	0	682.10	854.86
Professional Collaboration	0	768.51	947.08
Classroom Teaching Practice: Enhanced Activities	0	1032.14	1182.27
2013 Teacher Self-Reports			
Professional Collaboration	0	364.82	560.89
Efficacy in Instruction	0	409.57	587.87
Teacher-Student Relations	0	412.64	589.49
Effective Professional Development	0	529.07	704.82
Constructivist Belief	0	596.89	770.55

Scales	PPP	CI lower	CI higher
Exchange and Coordination for Teaching	0	601.01	799.07
Satisfaction with Current Work Environment	0	644.32	816.18
Classroom Disciplinary Climate	0	750.92	933.96
Efficacy in Classroom Management	0	850.55	1032.40
Efficacy in Student Engagement	0	1259.18	1435.02
Satisfaction with Profession	0	2542.49	2728.73
Needs for Professional Development in Subject Matter and Pedagogy	0	6484.66	6722.86
Need for Professional Development for Teaching for Diversity	0	10966.79	11234.03
Participation among Stakeholders	0	11400.22	11624.41

Notes:

1. The pair-wise difference in all loadings and intercepts is set as following a distribution of zero mean and 0.05 variance.
2. The model fit is inferred from the posterior predictive probability value (PPP) and the confidence intervals (CI). A model fits well and approximate invariance is supported if the PPP value is greater than zero and the CIs contains zero (van de Schoot, et al., 2013).

ANNEX 4: SUMMARY OF FIT INDEXES OF THE PAIR-WISE COMPARISONS (EACH CULTURE WITH THE INTERNATIONAL REFERENCE GROUP) OF THE *SATISFACTION WITH CURRENT WORK ENVIRONMENT* SCALE

Culture	Model	CFI	TLI	RMSEA
Abu Dhabi (UAE)	Configural	0.99	0.96	0.09
	Metric	0.99	0.98	0.07
	Scalar	0.98	0.98	0.07
Alberta (Canada)	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.07
	Scalar	0.98	0.98	0.07
Australia	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.07
	Scalar	0.98	0.98	0.06
Brazil	Configural	0.99	0.98	0.07
	Metric	0.99	0.98	0.06
	Scalar	0.96	0.95	0.1
Bulgaria	Configural	0.99	0.96	0.09
	Metric	0.98	0.97	0.08
	Scalar	0.98	0.98	0.07
Chile	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.07
	Scalar	0.96	0.96	0.09
Croatia	Configural	0.99	0.96	0.09
	Metric	0.99	0.98	0.07
	Scalar	0.97	0.97	0.09
Cyprus[9]	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.06
	Scalar	0.99	0.98	0.06
Czech	Configural	0.99	0.97	0.08
	Metric	0.98	0.97	0.08
	Scalar	0.95	0.94	0.12
Denmark	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.07
	Scalar	0.98	0.98	0.06
England (UK)	Configural	0.99	0.97	0.08

9. See note 7.

Culture	Model	CFI	TLI	RMSEA
	Metric	0.99	0.98	0.08
	Scalar	0.97	0.97	0.09
Estonia	Configural	0.99	0.96	0.09
	Metric	0.98	0.97	0.07
	Scalar	0.96	0.95	0.1
Finland	Configural	0.98	0.95	0.1
	Metric	0.98	0.97	0.08
	Scalar	0.98	0.97	0.08
Flanders (Belgium)	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.06
	Scalar	0.98	0.97	0.08
France	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.06
	Scalar	0.98	0.97	0.08
Georgia	Configural	0.99	0.97	0.07
	Metric	0.98	0.96	0.08
	Scalar	0.97	0.97	0.08
Iceland	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.06
	Scalar	0.99	0.99	0.05
Israel	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.06
	Scalar	0.98	0.98	0.07
Italy	Configural	0.99	0.97	0.07
	Metric	0.99	0.97	0.07
	Scalar	0.98	0.98	0.07
Japan	Configural	0.98	0.94	0.11
	Metric	0.98	0.97	0.08
	Scalar	0.97	0.97	0.08
Korea	Configural	0.99	0.96	0.1
	Metric	0.98	0.97	0.08
	Scalar	0.98	0.98	0.07
Latvia	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.07
	Scalar	0.97	0.97	0.08
Malaysia	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.06

Culture	Model	CFI	TLI	RMSEA
	Scalar	0.90	0.88	0.15
Mexico	Configural	0.99	0.96	0.08
	Metric	0.98	0.97	0.07
	Scalar	0.91	0.89	0.14
Netherlands	Configural	0.99	0.97	0.08
	Metric	0.99	0.97	0.07
	Scalar	0.98	0.98	0.07
New Zealand	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.07
	Scalar	0.98	0.98	0.07
Norway	Configural	0.99	0.98	0.07
	Metric	0.99	0.98	0.06
	Scalar	0.99	0.99	0.05
Poland	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.06
	Scalar	0.98	0.98	0.06
Portugal	Configural	0.99	0.98	0.07
	Metric	0.98	0.97	0.09
	Scalar	0.98	0.97	0.08
Romania	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.06
	Scalar	0.98	0.98	0.07
Russian Federation	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.07
	Scalar	0.98	0.97	0.07
Serbia	Configural	0.99	0.96	0.09
	Metric	0.99	0.98	0.07
	Scalar	0.98	0.97	0.07
Shanghai (China)	Configural	0.99	0.97	0.08
	Metric	0.98	0.96	0.09
	Scalar	0.97	0.96	0.09
Singapore	Configural	0.99	0.96	0.09
	Metric	0.99	0.98	0.07
	Scalar	0.97	0.97	0.08
Slovak Republic	Configural	0.99	0.97	0.08
	Metric	0.99	0.97	0.07
	Scalar	0.96	0.95	0.1

Culture	Model	CFI	TLI	RMSEA
Spain	Configural	0.99	0.96	0.1
	Metric	0.98	0.97	0.08
	Scalar	0.98	0.97	0.08
Sweden	Configural	0.99	0.97	0.07
	Metric	0.99	0.98	0.06
	Scalar	0.98	0.98	0.07
United States	Configural	0.99	0.97	0.08
	Metric	0.99	0.98	0.06
	Scalar	0.99	0.99	0.05

Notes:

1. The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.
2. The model fit was evaluated by the Tucker Lewis Index (TLI) (acceptable above 0.90), Comparative Fit Index (CFI) (acceptable above .90), and Root Mean Square Error of Approximation (RMSEA) (acceptable below 0.08).
3. The acceptance of a more constrained model is based on the change of CFI (acceptable within 0.01 from a less to a more constrained model) (Cheung and Rensvold 2002, Rutkowski and Svetina 2014).
4. The cultures that do not reach scalar invariance when comparing with the international average reference group have the country name bolded.