Comparing teachers' job satisfaction across countries: A multiple-pairwise measurement invariance approach

Abstract

There is much interest in comparing latent traits, such as teacher job satisfaction, in large international surveys. However, different countries respond to questionnaires in different languages and interpret the questions through different cultural lenses, raising doubts about the psychometric equivalence of the measurements. Making valid comparisons depends on the latent traits displaying scalar measurement invariance. Unfortunately, this condition is rarely met across many countries at once. Different approaches that maximise the utility of such surveys, but remain faithful to the principles of measurement invariance testing, are therefore needed. This paper illustrates one such approach, involving multiple-pairwise comparisons. This enables us to compare teacher job satisfaction in England to 17 of the countries that participated in TALIS 2013. Teacher job satisfaction in England was as low, or lower, than all of the 17 comparable countries.

1. Introduction

Social surveys often include a series of related questions, designed to measure the same underlying latent construct. Respondents' answers to these questions are then typically combined in order to form a scale. For instance, in this paper we consider four questions about job satisfaction asked to a sample of teachers, with a 'job satisfaction' score then derived. Academics and policymakers wish to use these scale scores in different ways, such as being the dependent or explanatory variable in a regression model, or to compare average scale scores across groups (e.g. does teacher job satisfaction differ by country, gender, ethnicity or social class?). The primary focus of this paper is the latter. Using international studies such as the Organisation for Economic Co-Operation and Development (OECD) Teaching and International Learning Survey (TALIS), is it possible to make fair and legitimate cross-national comparisons of the derived questionnaire scales?

There are two main motivations for this paper. The first stems from the long and extensive literature recognising that such scales (and, indeed, the individual questions that form them) may not function equivalently across different groups (Meredith, 1964; Putnick & Bornstein, 2016). This could be due to differences in language, history, culture, interpretation or understanding (Bornstein, 1995), or any combination of the above. Great care is therefore needed before scores on such scales are compared, with it being vital that the measurement properties are thoroughly investigated first. If measurement invariance is not established first, then it is unclear whether differences in values reflect genuine differences in the construct across countries, or merely country-specific differences in the way people respond to certain questions (Steenkamp & Baumgartner, 1998).

In response to this issue, an extensive literature on testing for 'measurement invariance' (MI) has emerged (for a recent survey, see Millsap, 2012). Entire papers are often devoted to establishing the measurement properties of questionnaire scales, including checking for the comparability of these scales across different groups (e.g. Byrne, 1993; Koomen, Verschueren, van Schooten, Jak, & Pianta, 2012). Methods for testing MI are therefore now well established in the social science literature. Yet most applications of these methodologies have focused upon testing the comparability of scales across a relatively small number of groups. Much less research has considered how best to approach MI testing when the number of potential comparators is large, as is common in cross-national research using large-scale international databases.

An example of such a challenge comes from the TALIS 2013 study, a large-sample investigation of teachers drawn from 37 countries across the world. The TALIS survey is based around a teacher questionnaire, with a number of different scales designed to capture different aspects of the teaching profession (e.g. teachers' job satisfaction, professional development opportunities, and self-efficacy). However, when the comparability of these scales across countries was investigated by the survey organisers, 'scalar invariance' (the level of invariance required to compare average scores across nations) was not met. Indeed, out of the fifteen teacher scales tested for invariance in the TALIS 2013 data, none met their scalar invariance criteria. The TALIS technical report therefore clearly warns users that the scale scores derived cannot be directly compared across the participating countries (Desa, Gonzalez, & Mirazchiyski, 2014). This is unfortunate, as there is clear academic and policy interest in understanding (for instance) the countries that offer the best and worst working conditions for teachers, and in identifying those nations where teachers' job satisfaction is particularly high or low.

One of the most likely reasons why the OECD reached this conclusion is that they were testing whether the TALIS 2013 scales were fully comparable across every single participating country. In other words, if they had found scalar MI to hold, one would have been able to compare every single country against one another, and thus 'rank' each nation according to their average scale score. This was however always likely to be an unrealistic goal. With such a diverse group of countries included in the study, it was highly unlikely that fully comparable scales could have ever been produced.

More importantly, we argue that having such a scale is not really what individual countries are actually interested in. Rather, what policymakers often want to do is ‘benchmark’ their single country of interest against the widest possible group of fair comparators. For instance, education policymakers in England are likely to be most interested in how job satisfaction of teachers *in England* compares to teachers in other parts of the world. They will, on the other hand, have little interest in how teachers in Iceland compare to those Brazil, or how South Korea compares to Estonia, in this respect. Critically, establishing measurement invariance to address such questions is likely to be somewhat easier. That is, instead of trying to create a universal scale which allows one to compare every single country in the database, a more realistic approach may be to create a scale within a single nation of interest (e.g. England) and then use standard MI approaches to establish the comparator nations where a genuinely comparable scale can be constructed. We argue that such an approach is likely to better manage the trade-off between ensuring comparisons are fair and meaningful, and addressing research issues of greatest national interest.

This paper is therefore dedicated to illustrating such an approach. Specifically, in our application we attempt to benchmark teacher job satisfaction in our country of interest (England) against the widest possible set of nations where fair and legitimate comparisons can be made. Job satisfaction is “…a pleasurable or positive emotional state resulting from the appraisal of one’s job or job experiences” (Locke, 1976, p. 1304) and is of long-standing interest to policymakers (see, for example: Bowers, 1955; Butler, 1961), who face recurring problems with retention and shortages of qualified teachers (Dolton, 2006). The secondary motivation for this paper is therefore to provide comparisons of teacher job satisfaction in England with other countries, in order to better understand the state of the teaching profession in England. We do this by estimating several Multiple Group Confirmatory Factor Analysis (MGCFA) models to test for measurement invariance of the job satisfaction scale between England and every other country included in the TALIS dataset. This ‘multiple pairwise’ approach to MI testing allows us to thoroughly consider the countries we can legitimately ‘benchmark’ England against, and thereby directly addressing the research question of greatest interest to this particular country.

To preview our key findings, we establish that fully trustworthy comparisons of average job satisfaction scores can be made between England and nine other countries, with ‘reasonable’ comparisons possible to a further eight countries. Comparability between England and other Anglophone nations, along with some Scandinavian countries, is particularly good. We find that teacher job satisfaction in England was much lower ($d < -0.2$) than in 10 of the 17 countries where comparisons could be made, and somewhat lower ($0.2 < d < -0.1$) than in a further four of the 17. Three countries have similar levels of job satisfaction to England, with no country performing substantially worse. We therefore find that teacher job satisfaction in England in 2012 was as low, or lower, than in all comparable countries. The main contribution of the paper is demonstrate the multiple-pairwise approach to testing for measurement invariance, which can be used to make valid comparisons across wide groups of international comparators. The multiple-pairwise approach is therefore of general value in analysing large scale international assessment data, in which the traditional approaches to measurement invariance strongly constrain the insights that can be extracted from the data.

2. Data

The Teaching and Learning International Survey (TALIS) is a large-scale international survey designed to gain insight into the teaching profession. There is increasing policy interest in the importance of teachers for pupil learning (Schleicher, 2011; World Bank, 2013) and TALIS is the only international survey which focuses on the working conditions of school teachers. TALIS was first run in 2008, then again in 2013 and 2018. We use the most recent data available at the time of writing (2013). In the 37 participating countries, schools were randomly selected as the primary sampling unit, with a minimum of 20 teachers then chosen from within each school. Countries are required to achieve a sufficiently high response rate (75 percent of schools and 50 percent of teachers) for the sample to be considered representative of the teacher population. Almost all countries met this criteria, except for the United States where the results may be subject to some degree of non-response bias. We exclude Iceland from our analysis due to their data not being publicly available. Our focus is upon 'ISCED level 2' (i.e. lower secondary school) teachers, with a final sample size of 117,876 drawn from across 36 countries. Further details are provided in Table 1.

<< Table 1 >>

Teachers were asked four questions to elicit their job satisfaction in relation to their working conditions, with each using a four-point scale (strongly agree, agree, disagree, strongly disagree):

- [TT2G46C] I would like to change to another school if that were possible (reverse coded).
- [TT2G46E] I enjoy working at this school.
- [TT2G46G] I would recommend my school as good place to work.
- [TT2G46J] All in all, I am satisfied with my job.

The survey organisers (the OECD) constructed a satisfaction with the working environment scale based upon teachers' responses (variable 'TJSENVS' in the international database).

Two checks were conducted prior to the analysis. First, a confirmatory factor analysis (CFA) was performed for the job satisfaction scale in each country separately. Countries which did not show an acceptable fit according to either the CFI or RMSEA (Bulgaria, Denmark, Estonia, Finland, Japan, Korea, the Netherlands, Serbia, Singapore, Spain, Abu Dhabi (United Emirates) and Alberta (Canada)) were excluded from the analyses. Second, we tested the internal consistency of the job satisfaction scale in each country separately. This was done using McDonald´s Omega (Zinbarg, Revelle, Yovel, & Li, 2005) as a measure of internal consistency. We chose McDonald´s Omega over the more widespread Cronbach´s Alpha, as it outperforms Alpha and imposes fewer model assumptions (Dunn, Baguley, & Brunsden, 2014; Peters, 2014). Both Malaysia (ω=0.694) and

Georgia (ω=0.684) were excluded on the basis of having omega below 0.7. This leaves 22 of the 36 countries in the main analysis.

3. Methodology

Recall our aim is to compare average levels of teacher job satisfaction in England to other countries – but only where legitimate and meaningful comparisons can be made. This is not straightforward in a cross-national context, where differences in languages and cultures may lead to differences in how teachers interpret and respond to such questions. The most common approach for investigating the legitimacy of making such comparisons is via 'measurement invariance' testing using multi-group confirmatory factor analysis (MGCFA; Steenkamp & Baumgartner, 1998).

The intuition behind this approach, with reference to the job satisfaction scale ('TJSENVS'), is presented in Figure 1. Ovals depict the unobserved latent construct we are trying to measure, while rectangles refer to observed teacher responses to the four job satisfaction questions. Specifically, Q_w^x represents a single TALIS question w in country x. The value λ_w^x is known as a factor loading. These quantify the strength of the relationship between the latent trait ('TJSENVS') and question w, in country x. On the other hand τ_w^x is known as the 'threshold', and is essentially equivalent to the constant term in a regression model (with respect to the relationship between TJSENVS for question w in country x).

<<Figure 1>>

Figure 1 can also be represented using the following equation for each country:

$$Y_w^x = \tau_w^x + \lambda_w^x \cdot \theta^x + \epsilon,$$

Where:
Y_w^x = Observed responses to question w in country x.
λ_w^x = Factor loading quantifying the relationship between question w and the latent trait in country x.

τ_w^x = The threshold value for the relationship between question w and the latent trait in country x.
$\theta^x =$ The latent factor (job satisfaction) we are trying to measure in country x.
$\epsilon =$ Error term.

The factor loadings (λ_w^x) and thresholds (τ_w^x) are the main properties of the job satisfaction model, and the key parameters used to test for 'measurement invariance' (i.e. comparability of the TJSENVS scale) across countries. In essence, testing for measurement invariance involves putting ever more constraints upon the factor loadings (λ_w^x) and thresholds (τ_w^x), to test whether model fit declines.

An Overview of Configural, Metric and Scalar Invariance

The three levels of measurement invariance important for our analysis are configural (level 1), metric (level 2) and scalar (level 3) (see Meredith & Teresi, 2006; Rutkowski & Svetina, 2014). All three levels need to hold if meaningful cross-country comparisons of latent scale scores (such as the TJSENVS scale) are to be made (Vandenberg & Lance, 2000). The most basic level of measurement invariance (configural) requires the same set of questions to be associated with the latent trait across all groups (Horn & McArdle, 1992). With respect to job satisfaction in TALIS, this means all four job satisfaction questions should be associated with the over-arching TJSENVS scale within each country we wish to compare. Returning to Figure 1, if the loadings λ_A^1,λ_B^1, λ_C^1 and λ_D^1 are all unequal to zero in country 1 (e.g. England), we also require them to be unequal to zero in country 2 (e.g. Australia), country 3 (e.g. Japan) and any other country we wish to compare.

The second level of invariance (metric) is more restrictive, since it also requires that the factor loadings (λ) are equal across groups (Steenkamp & Baumgartner, 1998). In our application, this means that the strength of the relationship between our job satisfaction scale (TJSENVS) and each of the four individual questions (w) must be the same across countries. In terms of Figure 1, this means that $\lambda_A^{x_1} = \lambda_A^{x_2}, \lambda_B^{x_1} = \lambda_B^{x_2}, \lambda_C^{x_1} = \lambda_C^{x_2}$ and $\lambda_D^{x_1} = \lambda_D^{x_2}$, in order for 'metric invariance' to hold between country 1 and country 2. If this level of invariance is established, then it is widely accepted that one can use the teacher job satisfaction scale as an independent variable in a cross-country regression model, and that the estimated parameters could be fairly compared. However, establishing metric invariance alone does not allow one to legitimately compare country mean scores upon the constructed scale. It is not be possible to say that job satisfaction is higher in country 1 than country 2, based upon metric invariance alone.

In order for such stronger statements to be made, the third level of 'scalar' invariance must also hold. This additionally requires that all thresholds (τ) in all the groups we wish to compare are also equal (Meredith, 1993). Again returning to Figure 1, we now also need $\tau_A^{x_1} = \tau_A^{x_2}, \tau_B^{x_1} = \tau_B^{x_2}$, $\tau_C^{x_1} = \tau_C^{x_2}$ and $\tau_D^{x_1} = \tau_D^{x_2}$. Only if these constraints are satisfied are we able to legitimately say that job satisfaction in country 1 is better or worse than in country 2.

Testing which Level of Measurement Invariance Holds

One way to establish whether a model is sufficiently 'good' is to examine whether it 'fits' the empirical data reasonably well. As MGCFA models are computed using the empirical covariance matrix, a 'good fit' means that the theoretical covariance matrix of the model is very similar to the empirical covariance matrix of the data. By adding additional parameter constraints, higher levels of measurement invariance usually means that the model fits the data less well. The question, therefore, is how much worse are we willing to allow the model to fit the data when we add in additional constraints?

This is essentially how measurement invariance is tested in cross-national research. A series of sequential MGCFA models are estimated, each adding in additional restrictions upon the λ and τ parameters. Various 'fit indices' are then examined to check whether imposing the additional constraints means the model fits the data significantly worse (i.e. is the model becoming too inconsistent with the empirical data). If the fit to the data becomes too bad as additional constraints are added, we reject the hypothesis that the next level of measurement invariance holds.

Choice of Fit Indices

Although a simple χ^2 test is sometimes used to test model fit, this is highly sensitive to sample size (Chen, 2007; Cheung & Rensvold, 2002). Hence a number of alternatives have been developed, all of which compare (to some extent) the model's chi-squared (χ^2) statistic to its degrees of freedom (Hox & Bechger, 1998). We draw upon two such indices commonly used in the cross-national literature.

The first is the Comparative Fit Index (CFI; Bentler, 1990), which compares properties of the constrained invariance model to an unconstrained model. Specifically, Kenny (2015) defines the CFI as:

$$CFI = \frac{d(Null\ Model) - d(Proposed\ Model)}{d(Null\ Model)}$$

Where: d = Model χ^2 – model degrees of freedom.

The CFI is constrained to have a minimum of 0 and a maximum of 1, with higher values indicating better model fit. Note that when testing for invariance, the CFI helps us to consider the trade-off between worse model fit (a higher χ^2 statistic) versus the simplification of the model (having more degrees of freedom available), due to the additional constraints placed upon the τ and λ parameter constraints. This, in turn, helps us to judge whether the additional assumptions being made at higher levels of invariance testing really do hold (e.g. with respect to metric invariance, that making the assumption that the λ parameters are equal across countries is reasonable).

We use the CFI for two reasons. First, because it has been shown to be robust in testing for invariance (Cheung & Rensvold, 2002). Second, because the cut-off values for the statistic have been shown to be similar across levels of invariance (Chen, 2007), which simplifies the communication of our results. Despite these advantages, CFI also has limitations. In particular, CFI is a *relative* fit index, meaning that it tells us how much worse our model becomes when adding in additional constraints against some baseline (reference) model. A clear limitation of the CFI is therefore that, if the baseline (reference) model does not fit well to begin with, then it may be quite difficult to make it substantially worse by adding additional constraints. Consequently, it is possible that the CFI could indicate that a high level of invariance holds (e.g. scalar), even when the absolute fit of the model is rather poor. For this reason, and in line with advice in the literature (Kline, 2015), we do not rely exclusively on CFI.

The second fit index we use is the Root Mean Squared Error of Approximation (RMSEA – see Steiger & Lind, 1980). Kenny (2015) defines this as:

$$RMSEA = \frac{\sqrt{\chi^2 - df}}{\sqrt{df.(N-1)}}$$

Where:

χ^2 = The model χ^2 statistic.

df = Model degrees of freedom.

N = Sample size.

As with the CFI, the RMSEA is constrained to have a maximum of 1 and minimum of 0. Unlike the CFI however, it is a 'badness-of-fit' index, with lower figures indicating a better model fit. Note that, in contrast to the CFI, the RMSEA only uses parameters from the current model, and does not rely upon comparison to some null/baseline model. The RMSEA is therefore an *absolute* measure of model fit, complementing the relative measure provided by the CFI (Rigdon, 1996). Our other justification for using the RMSEA is that it provides a precise measure of model fit (Putnick & Bornstein, 2016).

When using these fit indices to test for the first level of invariance (configural) only the absolute value of these indices are considered. For the CFI, the 'null' model for the configural invariance test has all λ parameters set to the same constant, and only the thresholds estimated. The implied latent job satisfaction scale within this null model would simply be a linear composite of teachers' responses to the four TALIS job satisfaction questions. However, when moving on to testing the second (metric) and third (scalar) levels of invariance, it is *change* in model fit from the previous level that becomes the relevant quantity (i.e. only Δ_{CFI} and Δ_{RMSEA} are taken into account). Importantly, the metric and scalar tests involve consideration of whether the additional constraints lead to substantial deterioration in model fit *relative* to the initial configural model.

The Use of Cut-Off Values

Unfortunately, there are no golden rules as to what cut-off values should be used for the CFI and RMSEA indices. There are, however, some rules of thumb. When testing configural invariance, Browne & Cudeck (1993) suggest models with an RMSEA ≤ 0.05 have a good fit, values up to 0.10 indicate at least mediocre fit, while those above 0.10 should not be accepted. For the CFI, values above 0.95 are treated as indicating adequate fit (e.g. Schermelleh-Engel, Moosburger, & Müller, 2003; Schreiber et al., 2006). Then, when testing for metric and scalar invariance, the model fit should not deteriorate by more than -0.01 in CFI (Cheung & Rensvold, 2002) and 0.01 in RMSEA (Putnick & Bornstein, 2016). The OECD used these traditional cut-off values to test for measurement invariance in the TALIS 2013 study, and we therefore also use them within our analysis (further details provided below).

'Multiple Pairwise' Approach to Measurement Invariance

The standard way of applying the above approach to international datasets such as TALIS is to run a giant MGCFA including all countries in a single model. The three levels of invariance are then tested for all countries, with a decision then made for each level based upon a single CFI and RMSEA statistic. For instance, for metric invariance one would test the constraint that the λ parameters are equal across all of the 22 countries. This, of course, is highly unlikely to hold true. Hence, for most questionnaire scales included in international surveys such as TALIS and PISA, scalar invariance (which we need to hold in order to compare mean scores across countries) rarely holds. However, on the occasions that scalar invariance does using this approach, it encourages researchers to compare any two countries that they wish. For instance, it would be assumed England could be legitimately compared to as diverse places as Australia, Germany, Japan and Mexico.

This, however, is not how many national governments and researchers actually use such datasets. Rather than being able to compare job satisfaction across every single possible pairwise comparison (e.g. England to Spain, Luxemburg to Korea, Germany to Sweden), often we want to benchmark a single country of particular interest (e.g. England) to the largest possible group of comparators (e.g. England to Spain, England to France, England to Japan). We therefore employ a different approach. Following Asil and Brown (2016), this involves conducting a series of pairwise MGCFA models in which England is compared against each of the other participating countries, one-at-a-time.

Judging Comparability

Traditionally, invariance testing evaluates the three invariance levels in order, stopping when the fit indices no longer support the introduction of additional parameter constraints. However, there is no consensus on which specific fit indices should be used, or the cut-off values to be applied (Putnick & Bornstein, 2016). Moreover, different fit indices can lead one to reach different conclusions regarding the measurement invariance of a scale.

We therefore suggest a different approach be taken when applying our 'multiple pairwise comparison' methodology. Specifically, a series of MGCFA models for England (our country of interest) and every other country will be conducted, for each of the three invariance levels, regardless of the outcome of the preceding test. For instance, even if the criteria for metric invariance is not met when comparing England to another country (e.g. Japan), we still conduct the test for scalar invariance. Two fit indices (CFI and RMSEA) will then be examined for each of the three models. This gives us a total six criteria allowing us to consider whether measurement invariance between England and each of the other countries holds:

- Configural invariance: Absolute model fit. RMSEA≤ 0.10 and CFI≥ 0.95
- Metric invariance: Change in model fit. $\Delta_{RMSEA} \leq 0.01$ and $\Delta_{CFI} \geq -0.01$
- Scalar invariance: Change in model fit. $\Delta_{RMSEA} \leq 0.01$ and $\Delta_{CFI} \geq -0.01$

Using these six criteria, we judge the 'trustworthiness' of each pairwise comparison between England and every other country in terms of average job-satisfaction scale scores. The following three levels of 'trustworthiness' are then set (see Table 2 for a summary):

- *Trustworthy*. All six criteria are met. This is equivalent to scalar invariance being consistently met (using two separate fit indices) under the traditional 'hierarchical' measurement invariance approach.
- *Reasonable*. Both of the configural criteria are met, but one of the four remaining criteria are not. Our rationale for prioritising the configural criteria is that this essentially sets the baseline against which the higher levels of invariance are tested. It is hence vital that a good initial benchmark is set. Note that our 'reasonable' classification is equivalent to scalar invariance holding for at least one of our two fit indices under the traditional 'hierarchical' approach.
- *Unreliable*. All other countries are categorised as unreliable. Any country that fails both the RMSEA and CFI criteria at a given invariance level (e.g. failure to meet the metric threshold according to *both* the RMSEA and CFI) is classed as being an unreliable comparator. Failing on either RMSEA or CFI at both the metric and scalar level will also results in being classed as unreliable.

<< Table 2>>>

How to Present the Results?

A final consideration when using this approach is how to present the substantive results. Specifically, one wants to ensure that only the country of interest (England in our example) is contrasted with other countries, and that two non-England countries are never compared. For this reason, we eschew graphs and use a tabular presentation instead. More specifically, we calculate effect size difference in job satisfaction scale scores between England and each other country using Cohen’s d (Cohen, 1988). We then categories countries into five separate groups:

- ‘Much lower than England’ ($d < -0.2$)
- ‘Lower than England’ ($-0.2 \leq d < -0.1$)
- ‘About the same’ ($-0.1 \leq d \leq 0.1$)
- ‘Higher than England’ ($0.1 < d \leq 0.2$)
- ‘Much higher than England’($d > 0.2$).

Information will also be presented as to whether the difference between England and each comparator country is statistically significant at conventional levels.

Application to the TALIS Data

Two data preparation steps were taken prior to us applying this approach. First, to avoid estimation problems and assure meaningful parameter estimates, categories were collapsed if they had less than twenty responses. One category needed collapsing for one question in eight of the 35 countries. One category needed collapsing in two questions in a set of five countries. For each country other than England, we therefore created one dataset, with the relevant categories collapsed. For England, we created three version to enable the pairwise invariance testing against each of the other countries using the same data structure. Second, and subsequently, we conducted imputation of missing values. Although the amount of missing data was small (averaging around four percent per country) it did reach around ten percent in some instances (e.g. Abu Dhabi). We assume these data are Missing At Random (MAR), and implement multiple imputation with predictive mean matching (e.g. Rubin, 1987). We generate five imputed datasets, which should be sufficient given this low level of missingness (Cheema, 2014). Imputation was applied to each of three English datasets. For the measurement invariance testing, each country is matched with the corresponding English dataset. This ensures that the testing is based on the same parameters and a comparable construct.

Note that the observed TALIS questionnaire items use an ordinal four-point scale. Although it is common for applied researchers to apply linear factor analytic models to such data, the ordinal nature of the item-data means that the underlying assumption of multivariate normality is unlikely to hold, which necessitates a different approach (O'Connell, Goldstein, Rogers, & Peng, 2008). Throughout this paper we therefore recognise the categorical nature of the data, using a robust weight least squares (WLSMV) estimator with THETA parameterization in MPlus. This essentially fits an ordered probit model to the item-response data (Muthén, Muthén, & Asparouhov, 2015). Consequently, while we assume that the latent TJSENVS variable is normally distributed, the actual outcome data (i.e. teachers' responses to the job satisfaction questions) are treated as ordered-categorical.

A final important feature of the TALIS data for our analysis is the complex survey design. Throughout our analysis we apply the final teacher weights to adjust for design features in the survey sampling and for the relatively small amounts of teacher non-response. To account for the hierarchical nature of the data (teachers nested within schools) all standard errors are clustered at the school-level. Although alternative approaches to handling hierarchical data are available (e.g. estimation of a two-level factor model) the benefits of doing so (e.g. decomposing job satisfaction into school and teacher level variances) are not the focus of this paper.

All data analyses were conducted in R (R Core Team, 2017) using the mice (van Buuren & Groothuis-Oudshoorn, 2011) and the MplusAutomation (Hallquist & Wiley, 2017) packages. Mplus (Muthén & Muthén, 1998-2017) was used for computing the MGCFA models.

4. Results

Conducting the Analyses for the Trustworthiness Criteria

To begin, we consider the two criteria relate to configural invariance. Recall from section 3 that if this level of invariance fails (according to either the RMSEA or CFI criteria) then the trustworthiness of comparisons will be classed as unreliable. Figure 2 presents the results, with the left-hand panel reporting the RMSEA and right-hand panel the CFI.

<< Figure 2 >>

All countries included in this stage of the analysis meet the cut-off values for both CFI and RMSEA. The rank order of countries on the two fit statistics is very similar (Spearman's rank = 0.75).

<< Table 3 >>

The next step is to test each comparison for metric invariance, which is judged by *change* in the two indices from the configural model. These results are presented in Figure 3, with the change in the CFI plotted along the horizontal axis and change in the RMSEA along the vertical axis. Further details can be found in Table 3.

<< Figure 3 >>

There are three points to note. First, there is a strong cross-country correlation between the Δ_{CFI} und Δ_{RMSEA} (Spearman's rank = 0.96), suggesting that countries that fail one of the metric criteria are likely to also fail the other. Second, three countries sit in the top-left hand quadrant (Portugal, Chile and Mexico). These nations all passed the configural tests, but clearly fail both of the metric tests. According to the criteria set out in section 3, the trustworthiness of comparisons between England and these countries will be considered 'unreliable'. Third, 14 of the 21 comparators meet both of the metric criteria. Four countries fail according to the RMSEA but pass according to the CFI. No country passes according to the RMSEA, but then fails according to the CFI.

In the final step of the process, we test for scalar invariance. These results are presented in Figure 4, with cross shapes depicting countries that failed on one of the metric invariance criteria, and asterisks illustrating countries that failed on both of the metric criteria.

<< Figure 4 >>

There is a linear relationship between the two criteria (Δ_{CFI} and Δ_{RMSEA}) and the correlation is roughly the same as in the configural and metric tests (Spearman's rank = 0.75). Many countries pass the scalar invariance test under one fit index (e.g. Δ_{CFI}) but fail according to the other (e.g. Δ_{RMSEA}). Only one country (Shanghai-China) fails the scalar invariance test according to both the CFI and RMSEA (and thus is automatically assigned to the 'untrustworthy' group). One country (Mexico) fails the scalar test according to the CFI, but not the RMSEA. Four countries fail according to only the RMSEA (Croatia, Flanders, Norway and the Czech Republic), but not the CFI. A total of 13 countries manage to pass the scalar invariance test according to both fit indices. More than half of these (9) have passed all our other criteria thus far, meaning comparisons between England and these countries are deemed to be fully trustworthy. In contrast, six of the remaining countries in the bottom-right hand quadrant of Figure 4 have either failed one or two the metric criteria. These six countries will therefore be classed as 'untrustworthy' comparators. This serves as an important reminder as to how it is possible to pass the higher invariance levels, even when the fit indices show an unacceptable fit for the preceding levels.

Table 3 provides an overview of our invariance testing results. Fully trustworthy comparisons of average job satisfaction scores can be made between England and nine other countries. This includes all three English-speaking countries (Australia, New Zealand and the United States), along with several Eastern European nations (Poland, Romania, Russia and Slovakia). It also includes one Scandinavian country (Sweden), with comparisons to Norway also deemed to be reasonable. In contrast, it is clear that comparisons cannot be made between England and the East Asian nations: the results for Shanghai and Singapore have been classified as untrustworthy. A similar conclusion holds with respect to comparisons between England and the lower and middle income countries that participated in TALIS 2013. An unreliable rating has been assigned to Chile, Mexico and Portugal. Consequently, our approach seems to have identified some broad 'clusters' of countries with similar characteristics within our various 'trustworthiness' groups.

Benchmarking Teachers' Job Satisfaction in England Compared to Other Countries

Table 4 illustrates how average levels of teacher job satisfaction compares between England and other comparable countries. The three columns indicate: (a) the trustworthiness of the comparison, (b) whether the difference between England and each country is statistically significant and (c) an indication of the magnitude of the difference based upon effect sizes (see table notes and section 3). Note that our analysis only allows pairwise comparisons between countries and two non-England countries cannot be compared.

<< Table 4 >>

Of the nine countries where fully trustworthy comparisons can be made, six have levels of job satisfaction 'much higher' than in England (effect size difference > 0.2). Five countries where a 'reasonable' comparison can be made also have 'much higher' job satisfaction. The countries with 'much higher' job satisfaction include three English-speaking nations (Australia, New Zealand and United States) and four European countries (Sweden, Norway, Italy and Belgium). A further four countries are classified as having 'higher' job satisfaction than in England (effect size difference of between 0.1 and 0.2) within the trustworthy and reasonable groups. These are Poland, Russia, Croatia and France. Consequently, out of the 17 countries where 'trustworthy' or 'reasonable' comparisons can be made, 14 countries have higher levels of teacher job satisfaction than in England. Three countries have similar levels of job satisfaction (Slovakia, Czech Republic and Latvia) and no country has a has a lower level of job satisfaction than England. Together, Table 4 therefore provides strong evidence that teacher job satisfaction in England is as low or lower than in almost every other country where a robust comparison can be made.

5. Conclusion

Social surveys often contain a series of related questions designed to measure the same underlying characteristic or viewpoint of a respondent. However, due to variation in culture, language and social norms (amongst other factors), different groups may respond to these questions in different ways. A substantial literature on Measurement Invariance (MI) has therefore emerged, providing a now well-established methodology for testing the comparability of latent scale scores across different groups. Although standard approaches in this literature tend to work quite well when the number of groups being compared is quite small, establishing the scalar level of invariance has proven to be challenging in cross-national research, when the number of groups (countries) is often quite large (Desa et al., 2014). This problem can to some extent be attributed to the survey organisers' requirement to construct a 'one size fits all' scale, which can be compared across all countries participating in such studies. Yet such a goal is, in our view, unrealistic and very unlikely to be achieved (as previous analysis of the TALIS 2013 has shown). In any case, traditional approaches to MI do not directly address the real issue of interest to individual countries, which is typically how does their particular nation compare to elsewhere. Alternative approaches to benchmarking individual countries is therefore needed, maximising the utility of cross-national surveys to address research questions of interest, while also remaining faithful to the principle of conducting fair measurement invariance tests.

Following Asil & Brown (2016), we have adopted one such an approach in this paper, where our goal has been to benchmark average scores on a teacher job satisfaction scale in one particular country (England) against as many international comparators as possible. To do so, we have estimated a series of pairwise MGCFA models, each including England and one of the other comparator countries. A set of six measurement criteria have then been set, based upon standard

MI approaches and cut-off values, to judge the trustworthiness of each pairwise comparison. Our results indicate that fully trustworthy comparisons can be made between England and nine other countries, with reasonable comparisons possible to a further nine. This includes three other English-speaking countries that were included in the TALIS database and showed an acceptable CFA for the construct, along with two Scandinavian nations. We find strong evidence that teacher job satisfaction in England in 2013 was as low or lower than all 18 comparable countries. We have also highlighted how fair comparisons of this scale could not be made between England and the East Asian nations, despite this group of countries currently being of great political and policy interest in England (e.g., Jerrim & Vignoles, 2016).

These findings should, of course, be interpreted in light of the limitations of this study, and indeed of this particular methodological approach. First, this technique can only be used when one's goal is to benchmark a single country of interest against international comparators. Although we argue that this is typically the most important goal of national policymakers, and is likely to provide them with a more robust and meaningful analysis than current approaches to MI testing within the international comparative literature, caution needs to be taken when presenting results so that they are not misinterpreted (e.g., so that comparisons between countries where MI has not been established are not made). Second, as with all MGCFA approaches, the six measurement criteria we have set are based around 'cut-off' values. There are no golden rules as to the exact values these should take and, consequently, we have followed established rules-of-thumb. Nevertheless, it is important to note that some of the judgements made regarding the comparability of scales would change if even some minor adjustments to the cut-offs used are made. Third, we remind readers that we have collapsed certain categories in our analysis in order to ensure sufficient cell sizes for model estimation. Although we do not expect this to have major implications for our results, this remain an important limitation of our data.

Despite these limitations, we believe the approach used in this paper could have a wide range of applications. For instance, with growing numbers of participants in the OECD's Programme for International Student Assessment (PISA) study, drawing fair comparisons across the diverse set of nations is becoming ever more difficult. Indeed, there is growing scepticism that results from such studies cannot truly be compared across all of the countries that take part. We therefore believe that there is potential for the approach set out in this paper to establish a fairer group of countries for each nation to compare themselves against, both in terms of questionnaire responses and the PISA cognitive test scores. In doing so, we hope that this paper helps to stimulate greater use of cross-national resources amongst national governments and researchers, particularly with respect to benchmarking key aspects of their education systems against other nations, but only where such comparisons can be reliably made.

References

Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, *16*(1), 71-93.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bornstein, M. H. (1995). Form and function: Implications for studies of culture and human development. *Culture & Psychology*, *1*(1), 123-137.

Bowers, N. D. (1955). *The development and initial validation of an instrument designed to appraise certain aspects of teacher job satisfaction*. University of Minnesota.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136−162), Newbury Park, CA: Sage.

Butler, T. M. (1961). Satisfactions of beginning teachers. *The Clearing House: A journal of educational strategies, issues and ideas*, *36*(1), 11-13.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

Byrne, B. M. (1993). The Maslach Burnout Inventory: Testing for factorial validity and invariance across elementary, intermediate and secondary teachers. *Journal of Occupational and Organizational Psychology*, *66*(3), 197-212.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487-508.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack measurement invariance. *Structural Equation Modeling*, *14*(3), 464-504.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd edition. Hillsdale, NJ: Erlbaum.

Desa, D., Gonzalez, E., & Mirazchiyski, P. (2014). Construction of scales and indices. In Belanger, J., Normandeau, S., & Larrakoetxea, E. (Ed.), *TALIS 2013 technical report* (pp. 145-295). Paris: OECD.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399-412.

Dolton, P. J. (2006). Teacher supply. *Handbook of the Economics of Education*, *2*, 1079-1161.

Hallquist, M., & Wiley, J. (2017). *MplusAutomation: Automating Mplus model estimation and interpretation.* R package version 0.7. Retrieved from https://CRAN.R-project.org/package=MplusAutomation

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117-144.

Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling. *Family Science Review*, 11, 354-373.

Jerrim, J., & Vignoles, A. (2016). The link between East Asian 'mastery' teaching methods and English children's mathematics skills. *Economics of Education Review*, *50*, 29-44.

Kenny, D. (2015). Measuring model fit. Accessed 15/02/2018 from http://davidakenny.net/cm/fit.htm

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press.

Koomen, H. M., Verschueren, K., van Schooten, E., Jak, S., & Pianta, R. C. (2012). Validating the Student-Teacher Relationship Scale: Testing factor structure and measurement invariance across child gender and age in a Dutch sample. *Journal of School Psychology*, *50*(2), 215-234.

Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1297-1343). Chicago: Rand McNally.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika*, *29*(2), 177-185.

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

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11), 69-77.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. NY: Routledge.

Muthén, B., Muthén, L., & Asparouhov, T. (2015). *Estimator choices with categorical variables.* Retrieved from https://www.statmodel.com/download/EstimatorChoices.pdf

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide*. Eight Edition. Los Angeles, CA: Muthén & Muthén.

O'Connell, A. A., Goldstein, J., Rogers, H. J. & Peng, C.-Y. J. (2008). Multilevel logistic models for dichotomous and ordinal data. In A. A. O'Connell and B. McCoach (Ed.), *Multilevel analysis of educational data* (pp. 199-242). Charlotte, NC: Information Age Publishing Inc.

OECD [Organisation for Economic Co-operation and Development] (2014). *TALIS 2013 technical report*. OECD, Paris.

OECD (2014). *TALIS 2013 User guide*. OECD, Paris.

Peters, G. J. Y. (2014). The alpha and the omega of scale reliability and validity: Why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. *The European Health Psychologist*, 16, 56-69.

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71-90.

R Core Team (2017). *R: A language and environment for statistical computing* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(4), 369-379.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York, NY: John Wiley & Sons.

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

Schermelleh-Engel, K., Moosburger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research*, *8*(2), 23-74.

Schreiber, J. B., Stage, K. F., King, J., Nora, A., & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323-337.

Schleicher, A. (2011). *Building a High-Quality Teaching Profession: Lessons from around the World*. Paris, France: OECD Publishing.

Steenkamp, J. B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78-90.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70.

World Bank. (2013). *Achieving learning for all*. New York City, NY, USA: World Bank.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's Alpha, Revelle's Beta, McDonald's Omega: Their relations with each and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123-133.

Table 1: Sample Sizes in Each Country

Country	Abbreviation	Sample size
Australia	AUS	2,059
Brazil	BRA	14,291
Chile	CHL	1,676
Chinese Shanghai	CSH	3,925
Croatia	HRV	3,675
Czech Republic	CZE	3,219
England	ENG	2,496
Flanders (Belgium)	BFL	3,129
France	FRA	3,002
Israel	ISR	3,403
Italy	ITA	3,337
Latvia	LVA	2,126
Mexico	MEX	3,138
New Zealand	NZL	2,862
Norway	NOR	2,981
Poland	POL	3,858
Portugal	PRT	3,628
Romania	ROU	3,286
Russian Federation	RUS	3,972
Slovakia	SVK	3,493
Sweden	SWE	3,319
United States of America	USA	1,926

Notes: Figures refer to countries participating in the ISCED level 2 (lower primary school) component of TALIS 2013. Only countries included in the analysis are shown in the table.

Table 2. Criteria Used to Judge the Trustworthiness of Comparisons of Average Job Satisfaction Scale Scores Between England and Other Participating Countries

Trustworthy	All six criteria met
Reasonable	5 of 6 criteria met, including both configural criteria
Unreliable	All countries that are neither trustworthy or reasonable

Notes: The six criteria are as follows. (1) RMSEA≤ 0.10 for the configural model; (2) CFI≥ 0.95 for the configural model; (3) $\Delta_{RMSEA} \leq 0.01$ for the metric model; (4) $\Delta_{CFI} \geq -0.01$ for the metric model; (5) $\Delta_{RMSEA} \leq 0.01$ for the scalar model; (6) $\Delta_{CFI} \geq -0.01$ for the scalar model.

Table 3. Measurement Invariance Test Coefficients Across Countries

			Configural		**Metric**		**Scalar**	
		McDonald´s ω	**RMSEA**	**CFI**	Δ_{RMSEA}	Δ_{CFI}	Δ_{RMSEA}	Δ_{CFI}
Trustworthy	Australia	0.799	0.073	0.998	-0.031	0.001	-0.009	0
	Israel	0.810	0.081	0.998	-0.015	-0.001	0.006	-0.004
	New Zealand	0.841	0.083	0.998	-0.031	0.001	-0.013	-0.001
	Poland	0.789	0.090	0.996	-0.031	0.001	0.004	-0.004
	Romania	0.811	0.094	0.997	-0.034	0.001	0.001	-0.003
	Russia	0.774	0.085	0.996	-0.017	-0.001	0.001	-0.004
	Slovakia	0.720	0.092	0.996	-0.001	-0.004	-0.004	-0.006
	Sweden	0.762	0.073	0.998	0.005	-0.003	-0.006	-0.003
	USA	0.849	0.072	0.998	-0.018	0	-0.009	-0.01
Reasonable	Belgium	0.827	0.076	0.998	-0.017	0	**0.023**	-0.005
	Brazil	0.736	0.051	0.997	**0.013**	-0.004	-0.014	-0.002
	Czech Republic	0.816	0.094	0.997	-0.03	0	**0.035**	-0.01
	France	0.797	0.079	0.998	**0.014**	-0.004	-0.009	-0.003
	Croatia	0.804	0.098	0.996	-0.019	0	**0.015**	-0.008
	Italy	0.779	0.065	0.998	**0.039**	-0.005	0.004	-0.009
	Latvia	0.723	0.058	0.999	**0.014**	-0.002	0.001	-0.003
	Norway	0.789	0.057	0.999	-0.019	0	**0.017**	-0.003
Unreliable	Chile	0.726	0.06	0.999	**0.08**	**-0.013**	-0.024	-0.003
	Shanghai	0.759	0.097	0.996	-0.002	-0.003	**0.018**	**-0.014**
	Mexico	0.715	0.07	0.998	**0.068**	**-0.013**	-0.002	**-0.013**
	Portugal	0.787	0.054	0.999	**0.093**	**-0.011**	-0.032	-0.002

Notes: Invariance fit indices refer to estimates from a two-country MGCFA model, including England and each individual comparator country. McDonald´s Omega in England is 0.847 for no collapsed category, 0.844 for one collapsed category and 0.859 for two. Bold font with grey shading indicates values that fail to meet our cut-off criteria. The left-hand column provides the final classification of the comparability of the job satisfaction scale between each country and England.

Table 4. Comparison of Teacher Job Satisfaction in England With Other Countries

Comparability of scale to England	Country	Significantly different to England	Teachers job satisfaction compared to England
Trustworthy	Australia	***	Much higher than England
	Israel	***	Much higher than England
	New Zealand	***	Much higher than England
	Romania	***	Much higher than England
	Sweden	***	Much higher than England
	USA	***	Much higher than England
	Poland	***	Higher than England
	Russia	***	Higher than England
	Slovakia	-	About the same
Reasonable	Brazil	***	Much higher than England
	Flanders (Belgium)	***	Much higher than England
	Italy	***	Much higher than England
	Norway	***	Much higher than England
	Croatia	***	Higher than England
	France	***	Higher than England
	Latvia	-	About the same
	Czech Republic	*	About the same

Notes: See section 3 for our definition of the three 'comparability' groups (trustworthy, reasonable and unreliable). *, ** and *** indicate that the mean of the job satisfaction scale in that country is significantly lower than in England at the 10, 5 and 1 percent levels respectively. The final column refers to the difference in the (job satisfaction) scale score mean compared to England in terms of an effect size. 'Much higher'/'Much lower' than England refers to an effect size difference of at least 0.20. 'Higher/lower' than England refers to an effect size greater than 0.1 but less than 0.2. 'About the same' refers to an effect size difference of less than 0.1. Results for Portugal, Mexico, Chile and Shanghai not reported due to unreliability of the comparisons.

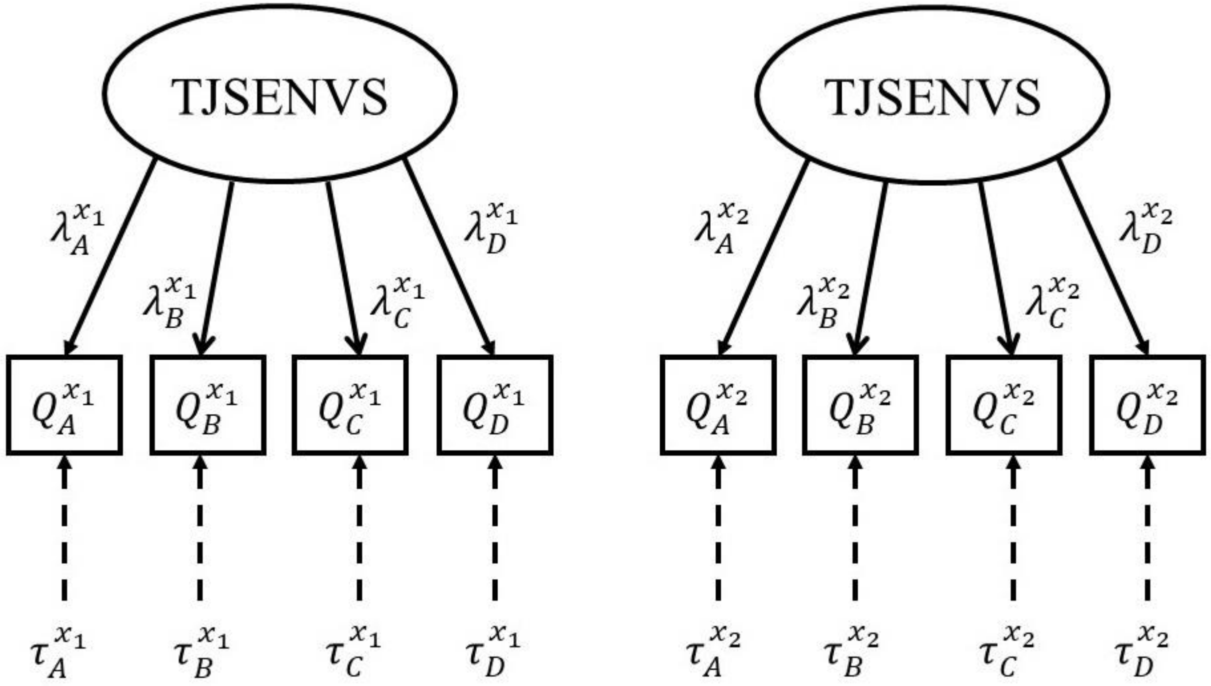

FIGURE 1. A hypothetical example of the MGCFA model to test invariance of the teacher job satisfaction scale (TJSENVS) across two countries.

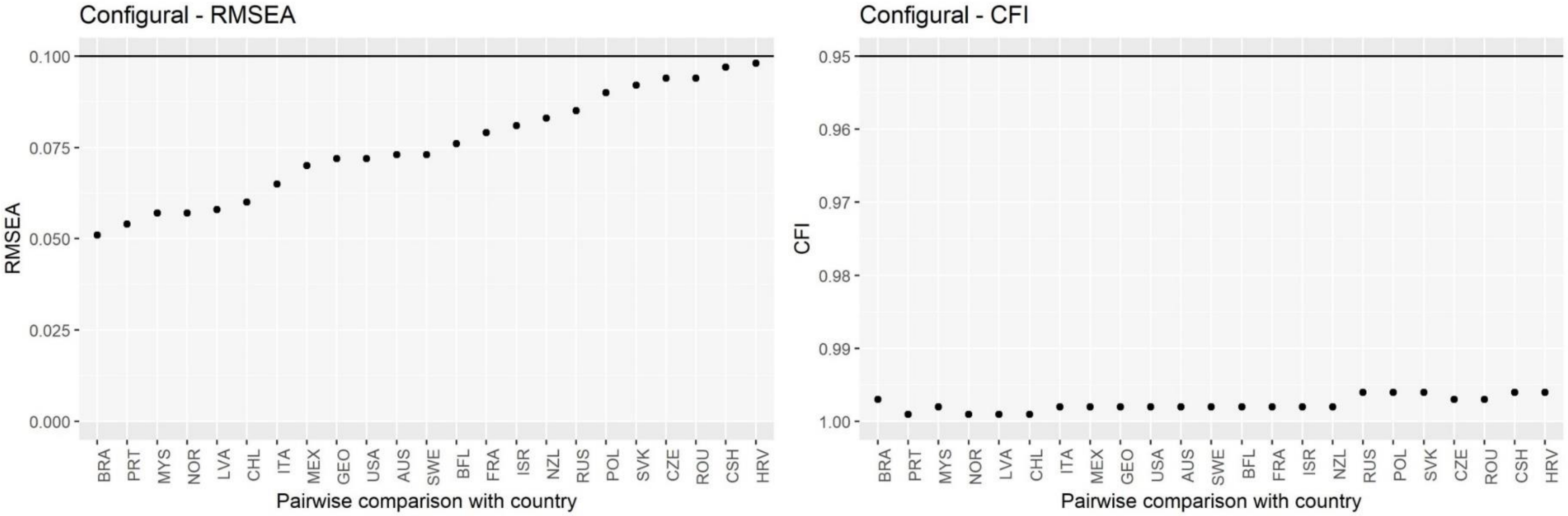

FIGURE 2. Results of pairwise tests for configural invariance between England and each comparator country. Points below the horizontal line illustrate where the criteria for configural invariance has been met.

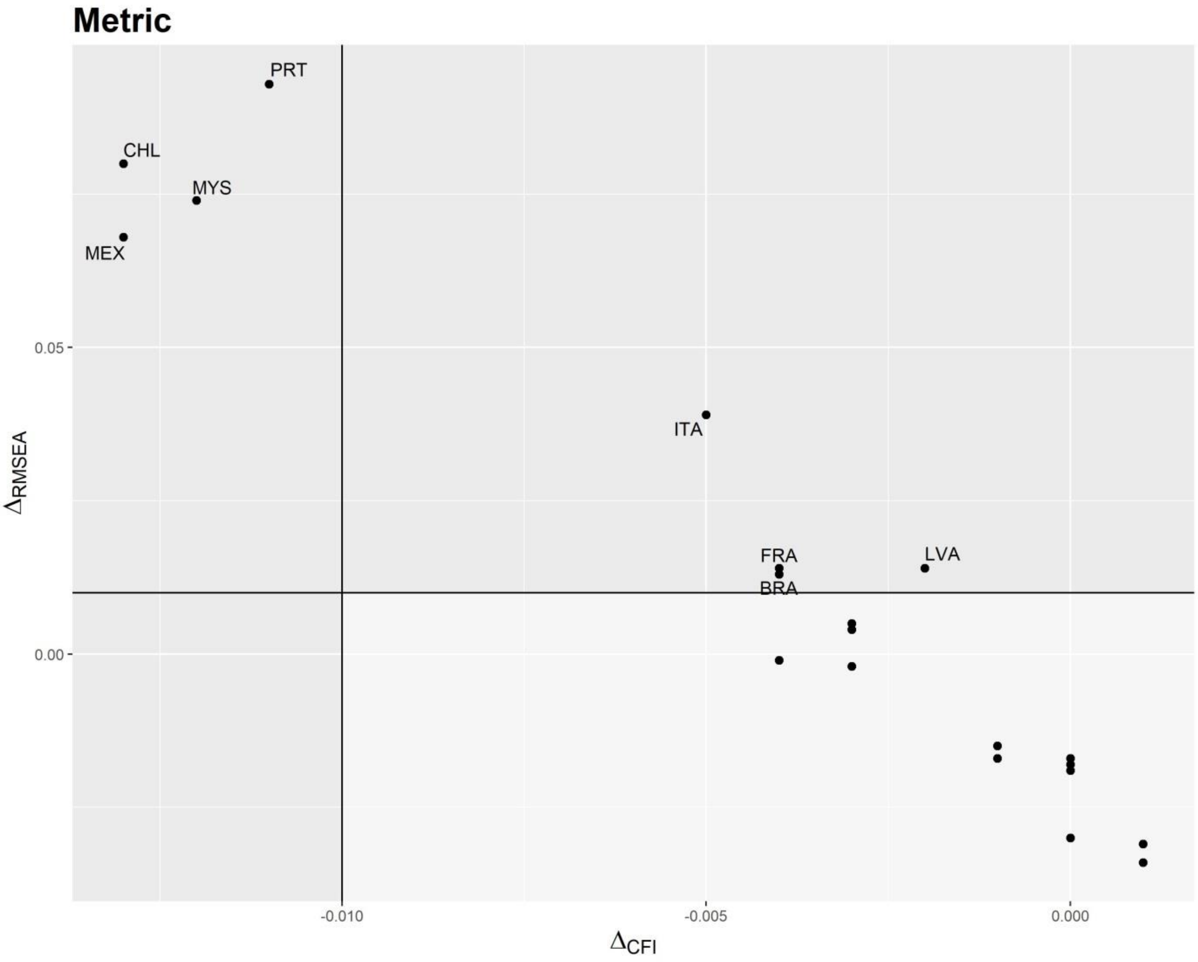

FIGURE 3. A comparison of Δ_{RMSEA} to Δ_{CFI} for the metric invariance tests. The bottom right hand corner illustrates where the criteria for both fit indices have been met.

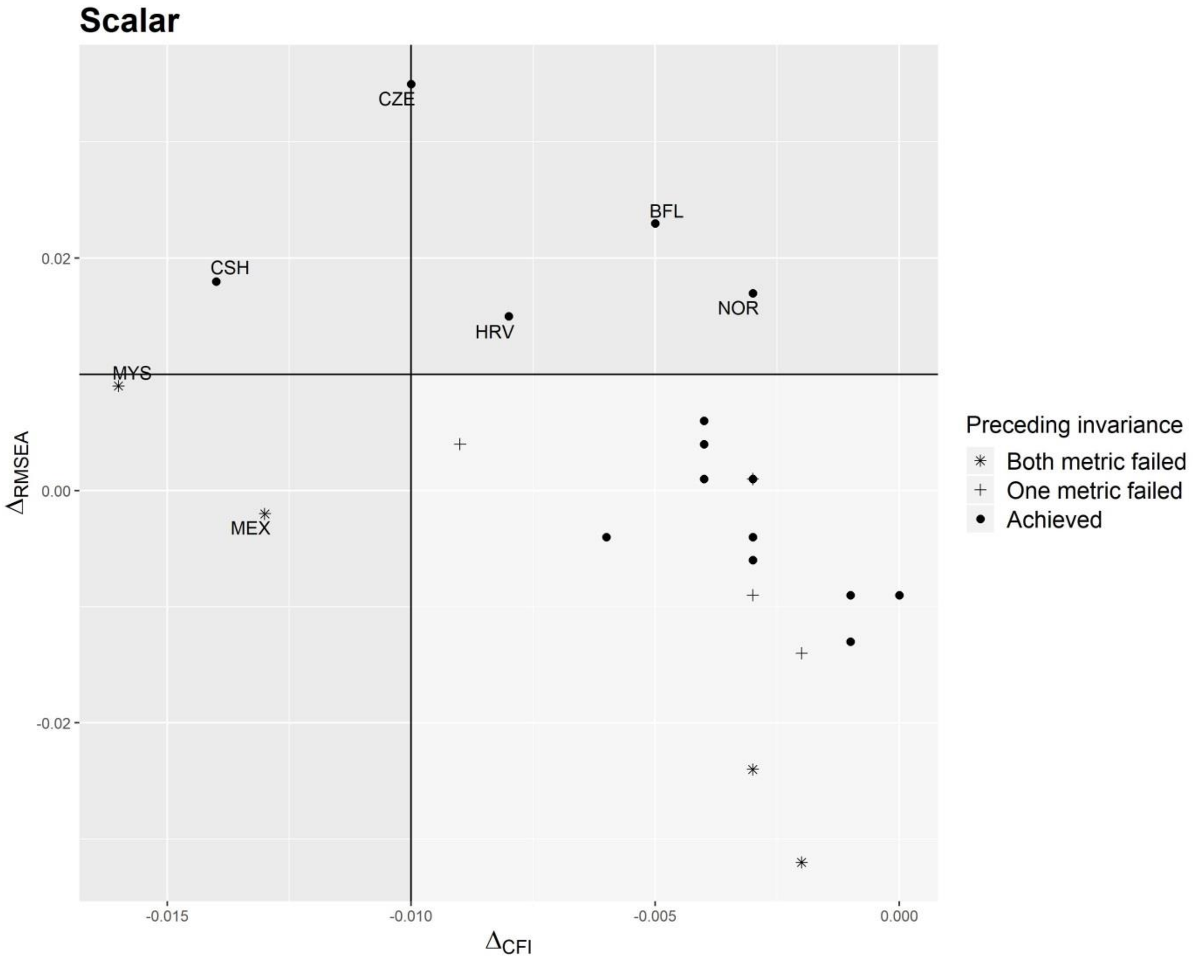

FIGURE 4. Fit statistics for scalar invariance tests. A comparison of Δ_{RMSEA} to Δ_{CFI} for the scalar invariance tests. The bottom right hand corner illustrates where the criteria for both fit indices have been met.