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

Laura Zieger[a], John Jerrim[b], Sam Sims[b,]

[a]Technical University of Dortmund

[b]UCL Institute of Education; Education Datalab

September 2018

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 22 of the 35 countries that participated in TALIS 2013. Teacher job satisfaction in England was as low, or lower, than all of the 22 comparable countries.

Key Words: TALIS; measurement invariance; job satisfaction, cross-national comparison.

Contact details: John Jerrim (J.Jerrim@ucl.ac.uk) Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way London, WC1H 0AL

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 and 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 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 et al., 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 et al., 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, however, was always likely to be an unrealistic and

unachievable 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. Teacher job satisfaction – defined as "…a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences" (Locke, 1976, p. 1304) - 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 13 (out of 35) potential comparator countries, with 'reasonable' comparisons possible to a further nine. Comparability between England and other Anglophone nations, along with most Scandinavian countries, is particularly good. We find that teacher job satisfaction in England was much lower ($d < -0.2$) than in 15 of the 22 countries where reasonable comparisons could be made, and somewhat lower ($0.2 < d < -0.1$) than in a further 5 of the 22. 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 2013 (TALIS 2013) is a large-scale international survey designed to gain insight into the teaching profession. 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 to strongly disagree):

- [TT2G46C] I would like to change to another school if that were possible[1].
- [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).

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

[1] We reverse code this item so all questions map in the same direction.

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.
$\lambda_w^x =$ 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. Basically, measurement invariance involves putting ever more constraints upon the factor loadings (λ_w^x) and thresholds (τ_w^x), to test whether three

hierarchical levels of invariance hold. These are configural (level 1), metric (level 2) and scalar (level 3). All three levels need to hold if meaningful cross-country comparisons of latent scale scores (such as the TJSENVS scale) are to be made (e.g. Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

An overview of configural, metric and scalar invariance

The most basic level of measurement invariance (configural) requires the same set of questions to be associated with the latent trait across all groups. 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 is more restrictive. It assumes that the factor loadings (λ) are equal across groups. In our application, this means that the strength of the relationship between our job satisfaction scale (TJSENVS) and each 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[2]. However, establishing metric invariance alone does not allow one to legitimately compare country mean scores upon the constructed scale; it

[2] For instance, one could model the relationship between teacher job satisfaction and teacher turnover (i.e. probability of leaving their job) in the two countries, and determine in which nation job satisfaction is the stronger predictor.

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. 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.

How do you test 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.

The 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 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).

Note that the CFI is a ***relative*** fit index; 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, we also use a second index, which provides an *absolute* measure of model fit – 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 CIF 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.

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 et al., 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).

Our '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 36 nations. 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, is 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 introduce a new approach to cross-national measurement invariance testing, involving a series of two-group MGCFA models, where England is compared against one of the other participating countries at a single time.

How do we judge which countries we can fairly compare England to?

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, along with Cronbach's Alpha as a measure of internal consistency. This

gives us a total seven criteria allowing us to consider whether measurement invariance between England and each other country holds:

- Internal consistency: Cronbach's Alpha > 0.70.
- 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 seven 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 four levels of 'trustworthiness' are then set (see Table 2 for a summary):

- *Trustworthy*. All seven 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 five 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[3]. 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.
- *Poor*. If either of the configural invariance criteria are not met, the cross-national comparability of the scale is deemed to be poor. Likewise, a poor rating is also assigned if two or more of our seven criteria are not met. At each invariance level (configural, metric, scalar) at least one of the RMSEA and CFI criteria must be passed. We do not believe these countries to be comparable to England, but we do present the results in order to illustrate the dangers of making such comparisons without first testing for invariance.
- *Unreliable*. Scales are automatically classified as unreliable if both RMSEA and CFI criteria are not met at any given invariance level (e.g. failure to meet the metric threshold

[3] If a poor initial benchmark is used, one may observed little or no deterioration when testing for metric and scalar invariance. However, the result is still a poorly fitting model.

according to ***both*** the RMSEA and CFI). An unreliable rating is also assigned if four or fewer of the seven criteria hold in total.

<< 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. 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 when one had less than twenty responses. In total, one category needed collapsing for one question[4] in eight of the 35 countries, with one category needing collapsing in two questions for five out of the 35 countries. Second, although the amount of missing data was small

[4] The question needing collapsing was J46 ("All in all, I am satisfied with my job") in the following countries: Bulgaria, Chile, Malaysia, Mexico, Norway, Portugal, Spain and Belgium.

(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) using five imputed datasets.

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 et al., 2008). Throughout this paper we therefore recognise the categorical nature of the data, using a robust weight least squares (WLSMV) estimator with THETA parameterization[5] in MPlus. This essentially fits an ordered probit model to the item-response data (Muthén et al., 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.

[5] THETA parameterization allows the residual variances of the latent trait to be parameters in the model, while excluding scale factors. For more information, see Mplus User's Guide (Muthén & Muthén, 1998-2017).

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.

3. Results

Conducting the analyses for the trustworthiness criteria

To begin, we consider the inter-rater consistency of the scale in each country using Cronbach's Alpha, with values greater than 0.7 deemed accepted. Only three countries failed to meet this criteria, Georgia ($\alpha = 0.661$), Malaysia ($\alpha = 0.643$) and Mexico ($\alpha = 0.676$), while England had one of the highest values ($\alpha = 0.837$). Further details can be found in Table 3. The TALIS job satisfaction scale therefore seems to have good levels of internal consistency in most TALIS countries.

Our next 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 considered either poor or unreliable. Figure 2 presents the results, with the left-hand panel reporting the RMSEA and right-hand panel the CFI.

<< Figure 2 >>

Eight of the 35 comparator countries sit above the 0.10 cut-off, including three East Asian nations (Japan, South Korea and Singapore), where there is clear evidence of a poor-fitting model. This group also includes Abu Dhabi and Finland, while Estonia, Serbia and Spain also just about miss the cut-off. Hence almost a quarter of potential comparator countries fails the RMSEA configural criteria. Interestingly, the same does not hold true in the right-hand panel for the CFI. Although the rank order of countries is similar (Spearman's rank = 0.89), each of

the 35 nations easily meets the 0.95 CFI cut-off (all values are actually very high, sitting around 0.99). In other words, according to this particular measure of fit, configural invariance is easily achieved for every single comparator country. This difference in conclusion is likely to be due to the different properties of these two measures; RMSEA is a measure of *absolute* fit while the CFI is a *relative* measure of fit (Rigdon, 1996). Consequently, it is possible for countries such as Japan and South Korea to reach configural invariance under a relative measure (CFI), but still have a poor fitting model overall and thus fail the RMSEA criteria. We believe this highlights how having clear rules based around more than one fit index (as per the criteria we set out in section 3) is particularly important when it comes to testing measurement invariance at the configural level.

<< 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. Orange dots highlight the countries where the configural invariance test was failed. Further details can be found in Table 3.

<< Figure 3 >>

There are four points to note. First, there is a strong cross-country correlation between the Δ_{CFI} und Δ_{RMSEA} (Spearman's rank = 0.96), indicating how countries that fail one of the metric criteria are disproportionately likely to also fail the other. Second, four countries sit in the top-left hand quadrant (Portugal, Malaysia, 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, most countries (26 out of 35) meet both of the metric criteria,

with five failing according to the RMSEA measure alone, along with the four countries that failed both criteria (as noted above). No country passes according to the RMSEA, but then fails according to the CFI. Finally, six of the seven countries that failed one of the configural tests passed ***both*** of the metric criteria (Spain is the exception, passing the metric test based upon the CFI but again narrowly failing on the RMSEA). This serves as an important reminder as to how it is possible for a comparison to pass the metric and scalar criteria, even when the fit of the initial configural model is rather poor. It is important to note that, under our criteria, comparisons between England and these six countries will be given a 'poor' trustworthiness rating, regardless of the fact that they have passed both metric criteria (due to failure to meet the RMSEA 0.10 threshold under the configural test). In contrast, a traditional hierarchal invariance testing approach based upon the CFI would at this point suggest that at least metric invariance for comparisons between England and these six countries may actually hold.

In the final step of the process, we test for scalar invariance. These results are presented in Figure 4, with orange markers depicting failure of configural invariance, and yellow/red markers illustrating failure of one/both of the metric criteria.

<< Figure 4 >>

Although there is still a linear relationship between the two criteria (Δ_{CFI} and Δ_{RMSEA}) as per the results for configural and metric invariance, the correlation is notable weaker (Spearman's rank = 0.57). It is therefore now reasonably common for countries to pass the scalar invariance test under one fit index (e.g. Δ_{CFI}) but fail according to the other (e.g. Δ_{RMSEA}). In-fact, only one country (Shanghai-China) fails the scalar invariance test according to both the CFI and RMSEA (and thus automatically assigned to the 'unacceptable' group). A further two countries (Mexico and Malaysia) fail the scalar test according to the CFI only, while four fail according to only the RMSEA (Estonia, Croatia, Flanders and the Czech Republic). Nevertheless, a total of 27 countries manage to pass the scalar invariance test using both fit indices. Around half of

these (13) have passed all our other criteria thus far, meaning comparisons between England and these countries are deemed to be fully trustworthy. In contrast, nine of the remaining countries in the bottom-right hand quadrant of Figure 4 have either failed to reach configural invariance or failed both the metric criteria. These nine countries will therefore fall within either the 'poor' or 'unacceptable' groups.

Table 3 provides an overview of our invariance testing results. Fully trustworthy comparisons of average job satisfaction scores can be made between England and 13 other countries. This includes all four English-speaking countries (Australia, Canada, New Zealand and the United States), along with several Eastern European nations (Bulgaria, Poland, Romania, Russia and Slovakia). It also includes two of the four Scandinavian countries (Denmark and 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, Japan, Korea, Malaysia and Singapore have all been classified as either poor or unacceptable. A similar conclusion holds with respect to comparisons between England and the lower and middle income countries that participated in TALIS 2013; a poor or unacceptable rating has been assigned to Abu Dhabi, Chile, Mexico, Malaysia and Serbia. 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 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 13 countries where fully trustworthy comparisons can be made, ten have levels of job satisfaction 'much higher' than in England (effect size difference > 0.2), along with an additional five countries where a 'reasonable' comparison can be made. This includes all four English-speaking nations (Australia, Canada, New Zealand and United States) and three of the four Scandinavian countries (Denmark, Sweden and Norway)[6]. Other notable countries where we can reasonably say teacher job satisfaction is higher than in England include European nations such as Italy, Belgium and the Netherlands. 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 reasonable/trustworthy groups. These are Poland, Russia, Croatia and France. Consequently, out of the 22 countries where 'trustworthy' or 'reasonable' comparisons can be made, 19 countries have higher levels of teacher job satisfaction than in England, three East European have similar levels (Slovakia, Czech Republic and Latvia) while no country has a has a lower level than England. Together, Table 4 therefore provides strong evidence that teacher job satisfaction in England is lower than in almost every other country where a robust comparison can be made.

There is a mix of results in the eight countries where comparisons to England are deemed to be of 'poor' quality. The three East Asian countries (Japan, Korea and Singapore) along with Estonia appear to have lower levels of teacher job satisfaction than in England, while in two European countries (Spain and Finland) and two middle-income countries (Serbia and Abu Dhabi) job satisfaction is higher. Note that we do not believe these comparisons to be valid and present the results here only to illustrate how misleading it is to can be to make these sorts of comparisons without first testing for invariance.

[6] Although teacher job satisfaction also appears to be a lot higher in Finland than in England, the quality of this particular comparison was considered poor.

4. 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 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.

We have adopted 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 seven 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 13 other countries, with reasonable

comparisons possible to a further nine. This includes all other English-speaking countries included in the TALIS database, along with three of the four Scandinavian nations. We find strong evidence that teacher job satisfaction in England in 2013 was as low or lower than all 22 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 miss-interpreted (e.g. so that comparisons between countries where MI has not been established are not made). Second, as with all MGCFA approaches, the seven 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. Finally, we have applied our approach to just one scale in a single cross-national database. It is important that other researchers now consider using such an approach in a wider range of cross-national analyses, particularly where 'benchmarking' a single participant is the primary goal.

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 underlying latent variables that are trying to capture) along with the actual 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

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.

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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 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. and Larrakoetxea, E. (Ed.), TALIS 2013 technical report (pp. 145-295), OECD.

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

George, D., & Mallery, P. (2003). *SPSS for Windows step by step: A simple style guide and reference, 11.0 update*, 4th edition. Boston: Allyn and Bacon.

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

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

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.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. 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* (prepared by IEA Data Processing and Research Center, Hamburg, Statistics Canada, Ottawa). OECD, Paris.

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.

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.

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

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In Bryant, K. J., Windle, M. E., & West, S. G. (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp.281-324).

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79-94.

Table 1: Sample sizes of participating TALIS country.

Country	Abbreviation	Sample size
Australia	AUS	2,059
Brazil	BRA	14,291
Bulgaria	BGR	2,975
Chile	CHL	1,676
Croatia	HRV	3,675
Czech Republic	CZE	3,219
Denmark	DNK	1,649
Estonia	EST	3,129
Finland	FIN	2,739
France	FRA	3,002
Georgia	GEO	2,759
Israel	ISR	3,403
Italy	ITA	3,337
Japan	JPN	3,484
Korea	KOR	2,933
Latvia	LVA	2,126
Malaysia	MYS	2,984
Mexico	MEX	3,138
Netherlands	NLD	1,912
New Zealand	NZL	2,862
Norway	NOR	2,981
Poland	POL	3,858
Portugal	PRT	3,628
Russian Federation	RUS	3,972
Serbia	SRB	3,857
Singapore	SGP	3,109
Slovakia	SVK	3,493
Spain	ESP	3,339
Sweden	SWE	3,319
England	ENG	2,496
United States of America	USA	1,926
Flanders (Belgium)	BFL	3,129
Abu Dhabi (United Emirates)	AAD	2,433
Alberta (Canada)	CAB	1,773
Romania	ROU	3,286
Chinese Shanghai	CSH	3,925

Notes: Figures refer to countries participating in the ISCED level 2 (lower primary school) component of TALIS 2013.

Table 2. Criteria used to judge the trustworthiness of comparisons of average job satisfaction scale scores between England and other participating countries

Trustworthy	All seven criteria met
Reasonable	6 of 7 criteria met, including both configural criteria
Poor	5 of 7 criteria met and at least one of CFI or RMSEA criteria are met at each invariance level
Unreliable	Both CFI and RMSEA criteria failed at an invariance level or 4 or fewer criteria met in total

Notes: The seven criteria are as follows. (1) Cronbach's Alpha > 0.70; (2) RMSEA≤ 0.10 for the configural model; (3) CFI≥ 0.95 for the configural model; (4) $\Delta_{RMSEA} \leq 0.01$ for the metric model; (5) $\Delta_{CFI} \geq -0.01$ for the metric model; (6) $\Delta_{RMSEA} \leq 0.01$ for the scalar model; (7) $\Delta_{CFI} \geq -0.01$ for the scalar model.

Table 3. Measurement invariance test coefficients across countries

		Cronbach's α	Configural		Metric		Scalar	
			RMSEA	CFI	Δ_{RMSEA}	Δ_{CFI}	Δ_{RMSEA}	Δ_{CFI}
Trustworthy	Australia	0.782	0.073	0.998	-0.031	0.001	-0.009	0
	Bulgaria	0.753	0.096	0.996	-0.003	-0.003	-0.006	-0.004
	Canada	0.803	0.090	0.998	-0.031	0	-0.012	0
	Denmark	0.831	0.091	0.997	-0.027	0.001	0.003	-0.002
	Israel	0.801	0.081	0.998	-0.015	-0.001	0.006	-0.004
	Netherlands	0.787	0.089	0.997	-0.034	0.001	0	-0.001
	New Zealand	0.828	0.083	0.998	-0.031	0.001	-0.013	-0.001
	Poland	0.779	0.090	0.996	-0.031	0.001	0.004	-0.004
	Romania	0.799	0.094	0.997	-0.034	0.001	0.001	-0.003
	Russia	0.762	0.085	0.996	-0.017	-0.001	0.001	-0.004
	Slovakia	0.71	0.092	0.996	-0.001	-0.004	-0.004	-0.006
	Sweden	0.745	0.073	0.998	0.005	-0.003	-0.006	-0.003
	USA	0.839	0.072	0.998	-0.018	0	-0.009	-0.01
Reasonable	Belgium	0.816	0.076	0.998	-0.017	0	**0.023**	-0.005
	Brazil	0.716	0.051	0.997	**0.013**	-0.004	-0.014	-0.002
	Czech Republic	0.808	0.094	0.997	-0.03	0	**0.035**	-0.01
	France	0.755	0.079	0.998	**0.014**	-0.004	-0.009	-0.003
	Georgia	**0.661**	0.072	0.998	0.004	-0.003	-0.004	-0.003
	Croatia	0.795	0.098	0.996	-0.019	0	**0.015**	-0.008
	Italy	0.759	0.065	0.998	**0.039**	-0.005	0.004	-0.009
	Latvia	0.703	0.058	0.999	**0.014**	-0.002	0.001	-0.003
	Norway	0.777	0.057	0.999	-0.019	0	**0.017**	-0.003
Poor	Abu Dhabi	0.736	**0.126**	0.993	-0.046	0.002	0.006	-0.006
	Spain	0.744	**0.103**	0.996	**0.011**	-0.006	-0.005	-0.006
	Estonia	0.734	**0.105**	0.995	-0.022	0	**0.018**	-0.01
	Finland	0.783	**0.120**	0.995	-0.031	0	-0.007	-0.002
	Japan	0.756	**0.152**	0.989	-0.06	0.004	-0.008	-0.004
	Korea	0.774	**0.169**	0.989	-0.058	0.002	-0.005	-0.007
	Singapore	0.794	**0.120**	0.995	-0.046	0.002	-0.002	-0.003
	Serbia	0.764	**0.103**	0.995	-0.032	0.001	0.001	-0.004
Unreliable	Chile	0.712	0.06	0.999	**0.08**	**-0.013**	-0.024	-0.003
	Shanghai	0.747	0.097	0.996	-0.002	-0.003	**0.018**	**-0.014**
	Mexico	**0.676**	0.07	0.998	**0.068**	**-0.013**	-0.002	**-0.013**
	Malaysia	**0.643**	0.057	0.998	**0.074**	**-0.012**	0.009	**-0.016**
	Portugal	0.76	0.054	0.999	**0.093**	**-0.011**	-0.032	-0.002

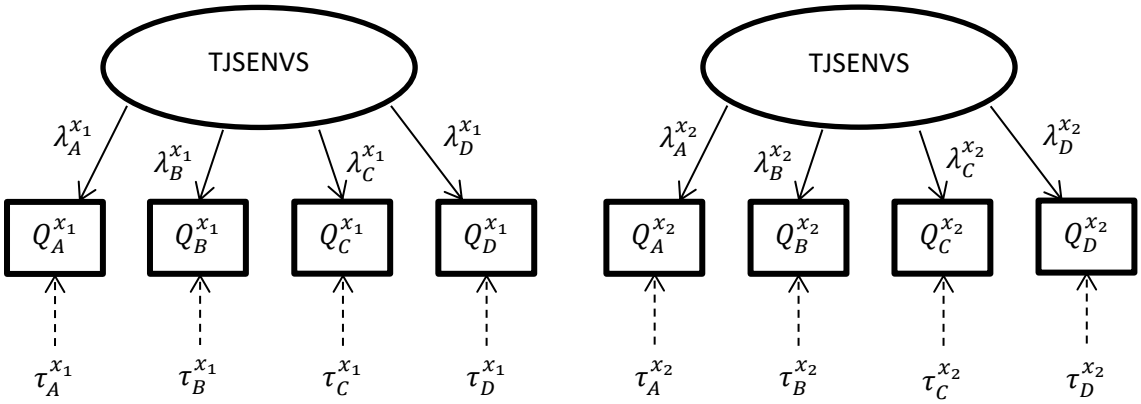
Notes: Invariance fit indices refer to estimates from a two-country MGCFA model, including England and each individual comparator country. Cronbach's Alpha in England is 0.837. Bold font with grey shading indicates values that fail to meet our cut-off criteria. The right-hand column provides the final classification of the comparability of the job satisfaction scale between each country and England.

Table 4. A comparison of teacher job satisfaction in England against other countries

Comparability of scale to England	Country	Significantly different to England	Teachers job satisfaction compared to England
Trustworthy	Australia	***	Much higher than England
	Bulgaria	***	Much higher than England
	Canada Alberta	***	Much higher than England
	Denmark	***	Much higher than England
	Israel	***	Much higher than England
	Netherlands	***	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
	Georgia	***	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
Poor	Finland	***	Much higher than England
	Spain	***	Much higher than England
	Abu Dhabi	***	Higher than England
	Serbia	***	Higher than England
	Estonia	***	Lower than England
	Japan	***	Much lower than England
	Korea	***	Much lower than England
	Singapore	***	Much lower than England

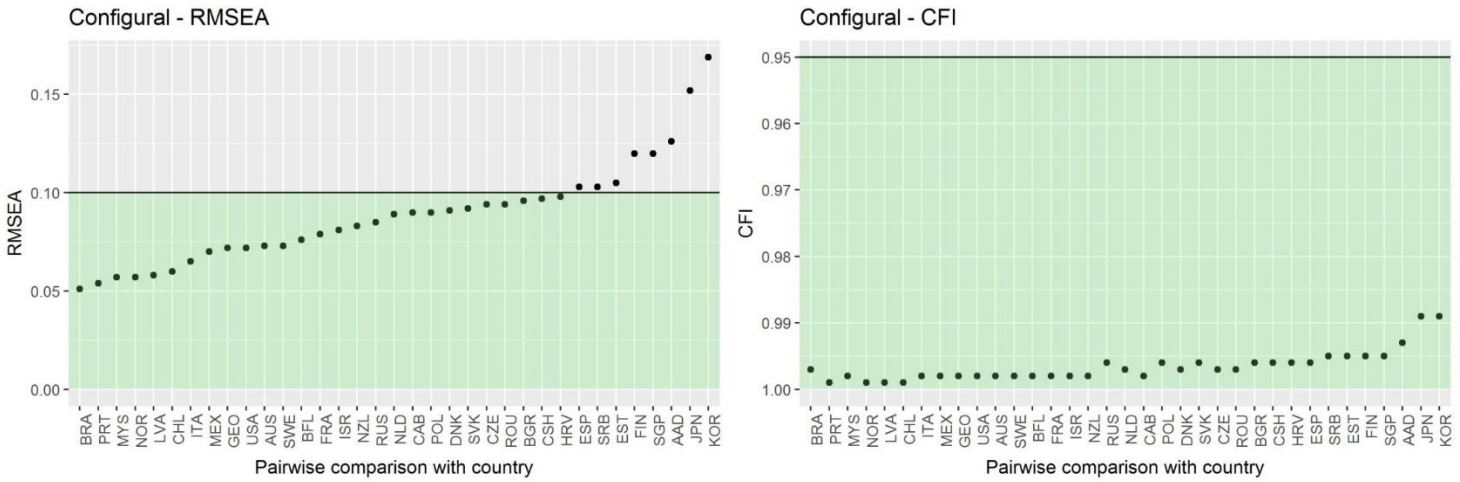
Notes: See section 3 for our definition of the four 'comparability' groups (trustworthy, fair, poor and unreliable). *, ** and *** indicate that the mean of the job satisfaction scale in that country is significantly lower than in England at the 10, 5 and 1 percent levels respectively. The final column 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, Malaysia, 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 across two countries.



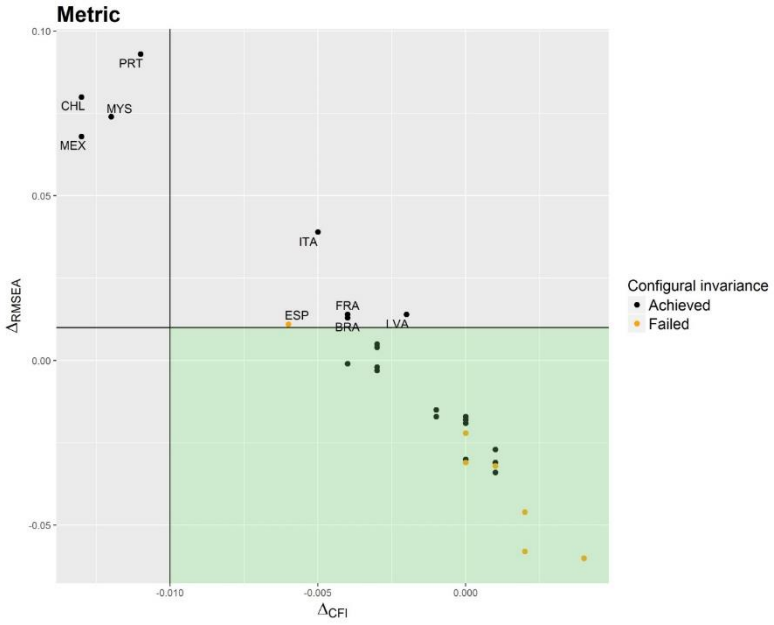
Notes: TJSENVS refers to the teacher job satisfaction latent variable.

Figure 2. Results of pairwise tests for configural invariance between England and each comparator country.



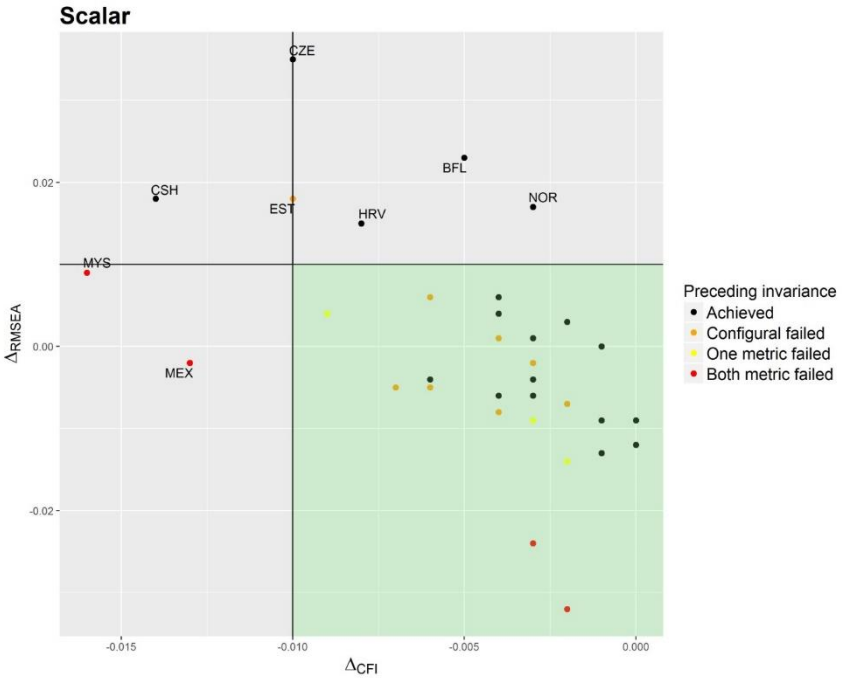
Notes: Left-hand panel presents results for the RMSEA and the right-hand panel for the CFI. Points below the horizontal line (within the green shaded area) illustrate where the criteria for configural invariance has been met. See Table 1 for country abbreviations.

Figure 3. Fit statistics for metric invariance tests. A comparison of Δ_{RMSEA} to Δ_{CFI}



Notes: Vertical axis presents results for the change in the RMSEA between the configural and metric MGCFA models. Horizontal axis provides the analogous results for the CFI. Shaded green area in the bottom right hand corner illustrates where the criteria for both fit indices have been met. Cross-reference with Table 3 for further details.

Figure 4. Fit statistics for scalar invariance tests. A comparison of Δ_{RMSEA} to Δ_{CFI}



Notes: Vertical axis presents results for the change in the RMSEA between the metric and scalar MGCFA models. Horizontal axis provides the analogous results for the CFI. Shaded green area in the bottom right hand corner illustrates where the criteria for both fit indices have been met. Cross-reference with Table 3 for further details.