Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance

Mustafa Asil

The University of Otago

Gavin T L Brown

The University of Auckland

Submission Date: 12/06/2015

Published as:
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. doi: 10.1080/15305058.2015.1064431

Abstract

The use of PISA across nations, cultures, and languages has been criticized. The key criticisms point to the linguistic and cultural biases potentially underlying the design of reading comprehension tests, raising doubts about the legitimacy of comparisons across economies. Our research focused on the type and magnitude of invariance or non-invariance in the PISA Reading Comprehension test by language, culture, and economic development relative to performance of the Australian English speaking reference group used to develop the tests. MG-CFA based on means and covariance structure (MACS) modeling was used to establish a d_{MACS} effect size index for each economy for the degree of non-invariance. Only three wealthy, English-speaking countries had scalar invariance with Australia. Moderate or large effects were observed in just 31% of the comparisons. PISA index of economic, social and cultural status (ESCS) had a moderate inverse correlation with dMACS suggesting that socio-economic resourcing of education played a significant role in MI, while educational practice and language factors seemed to play a further small role in non-invariance. Alternative approaches to reporting PISA results consistent with non-invariance are suggested.

Keywords: Measurement invariance (MI), Programme for International Student Assessment (PISA), Reading literacy

Interest in the quality of comparative studies conducted under the auspices of the Organization for Economic Co-Operation and Development (OECD) (i.e., the Programme for International Student Assessment, PISA) and the International Association for the Evaluation of Educational Achievement (IEA) (e.g., Trends in International Mathematics and Science Study –TIMSS; the Progress in International Reading Literacy Survey-PIRLS) is growing. These multilingual and multicultural assessment systems tend to develop tests in one language and adapt (including translation) them for other languages and countries. While these international tests systems have had a powerful impact on many countries, their legitimacy depends on the validity of test score comparisons across countries.

Test scores obtained from adapted and/or translated tests cannot be assumed to be comparable (AERA, APA, & NCME, 1999; ITC, 2000) unless scalar invariance across languages is present (Ercikan & Lyons-Thomas, 2013; ITC, 2000; Schmitt & Kuljanin, 2008). There are reasonable grounds to doubt the invariance of responses to tests despite careful adaptation processes. For example, recent studies (Boroditsky, 2001, 2011; Boroditsky, Fuhrman & McCormick, 2010; Boroditsky & Gaby, 2010; Fausey & Boroditsky, 2011) have shown that speakers of different languages differ in how well they can remember who did what, how they describe events, and how they think of time and space. Furthermore, a valid common instrument that assesses people from different languages and cultures may not even be feasible because of the complex nature of factors influencing reading comprehension skills (e.g., language, culture, cognitive, and economic development).

Even though many equivalence studies have been carried out at the item level (e.g., differential item functioning) and at the scale level (i.e., measurement invariance/equivalence) to ensure equivalence of international survey tests across cultures or languages, it is noteworthy that most of these studies have been conducted in mathematics and science domains. Reading literacy is likely to be greatly influenced by the linguistic and cultural features of the assessed countries (Grisay & Monseur, 2007). Hence, there is a need to confirm whether a reading test created in one language or jurisdiction can equally assess reading in other languages and societies. Therefore, in this study we evaluate the level of invariance in PISA 2009 reading comprehension across countries by considering unavoidable language and cultural differences.

Language, Culture, and Reading

Reading literacy as implemented by OECD involves "understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2010a, p.23). Reading literacy performance is influenced by a number of characteristics, such as the nature of each language, the type of writing system used to enact literacy, culturally-defined styles and approaches to teaching and learning, and levels of socio-economic development and investment in education. By evaluating post-facto whether patterns of invariance or non-invariance in test scores can be detected, these factors may give partial explanations for non-invariance in test scores.

The historic origins and linguistic characteristics of language families are well established (e.g., Fromkin & Rodman, 1978; McWhorter, 2003). For example, Western European languages fall predominantly into two family groups; that is, Germanic (including English, German, Dutch, etc.) and Italic (e.g., French, Spanish, etc.), both of which are part of a larger linguistic group derived from Proto-Indo European covering languages such as diverse as Hindi, Russian, Greek, and Gaelic. In contrast, languages in East Asia (e.g., Mandarin, Cantonese, and Japanese) are historically related to each other in structure, semantics, and syntax. Furthermore, thanks to European colonization processes in the 17th to 19th centuries, many nations use as a medium of instruction, a language from Western Europe (predominantly English, French, or Spanish). Thus, it may be that reading performance will be similar among countries that use the same language or whose languages belong to the same language family.

The writing system of language also impacts on reading performance (Perfetti & Harris, 2013). Three writing systems exist; (1) alphabetic in which written symbols correspond to spoken sounds (e.g., English, German, Korean), (2) logography (i.e., Chinese) in which written characters correspond to words, and (3) Japanese Kana syllabary in which written syllables map onto spoken syllables. Despite differences in writing, reading activates both meaning and phonological systems of language so that, despite differences in scripts, once mastery of the written system is achieved, reading in a language becomes efficient. This suggests that for 15-year olds in PISA reading comprehension, students should have mastered reading in the script and language being used to test reading. Thus, we should not expect to see that the type of script being read interferes with reading efficiency or invariance of reading scores.

A second major component in evaluating differences in reading performance has to do with the differences between cultures in the kinds of values and practices they prioritise within education. Hofstede (2007), based on years of inter-country comparisons, indicated culture is a collective way of thinking and behaving that distinguishes one group of people from another. Important characteristics of group differences have to do with (a) the distance between authority figures and subordinates (Power Distance), (b) the relative emphasis on the group or the individual as the source of authority (Collectivism vs. Individualism), (c) degree to which uncertainty is tolerated (Uncertainty Avoidance), (d) the tendency to prioritise stereotypically male or female values (e.g., assertiveness vs. service) (Masculinity vs. Femininity), and (e) tendency to focus on the future as opposed to the present or past (Long-term vs. Short-term Orientation) (Hofstede, 2007). An important caution is that, even when societies have shared histories, there are strong cross-national differences in these important characteristics (e.g., Japan and China diverge on uncertainty avoidance despite being similar on collectivism and power distance).

Cultural attributes are manifest in educational practices such as how teaching is understood or how assessment is conducted. In terms of educational practices, cross-cultural psychology research has shown that the values of high power distance between authorities and subordinates and high collectivism in Confucian-Heritage societies (e.g., China, Hong Kong, Taiwan, Korea, Singapore, and Japan) manifest themselves in high respect for the teacher and/or test and conformity to the group's goals and values over one's own (Hofstede & Bond, 1988). These conditions contribute to strong performance on achievement measures since such performance has important meaning to the student (e.g., fulfilling duty to the family—Peterson, Brown, & Hamilton, 2013). These societies, being much less affected by individualistic priorities, tend to emphasise didactic teaching practices that focus on transmitting knowledge that is evaluated in high-stakes public examinations (Shuell, 1996). In contrast, Western cultures, being defined by much lower power distance and stronger emphasis on the individual, promulgate school systems characterized by an emphasis on teachers attentively customizing teaching and curriculum practices to meet the needs and preferences of each individual child (Stobart, 2006) and in which tested performance is not the sole arbiter of a child's value or importance. This child-centered approach is most prevalent in economically-developed English speaking countries of the British Commonwealth (e.g., UK, New Zealand, Canada, and Australia). Hence, the degree to which cultural values, as evidenced by educational practices, are similar may be a confounding factor in reading performance.

A third factor that distinguishes countries is the level of funding and socio-economic resources dedicated to the educational system. Hofstede (2007) makes it clear that economic development tends to independent of the cultural characteristics he discusses. Within Asian societies, families commit significant proportions of household income to school fees and extra-curricular tuition (Brown & Wang, 2013). In contrast, Western developed economies expend significant proportions of GDP raised by taxation from the public economy on teacher salaries and school infrastructure (OECD average is about $10,000 USD per pupil; OECD, 2013) meaning that, on average, relatively little is spent additionally by most families. Societies that make significant public investment in education and which are highly developed may stimulate higher levels of reading performance than those which have relatively limited public expenditure on education.

Policy Impact of PISA

PISA focuses on the application of knowledge in mathematics, science, and reading to problems confronted in real-life situations and provides an international benchmark for participating countries. Policy makers and educators are taking PISA outcomes very seriously and many countries/economies are making important changes in their educational systems (Breakspear, 2012) and the influence of PISA is increasing over time (OECD, 2010b). Responses and policy reactions to PISA assessments, however, differ from country to country (Baird et al., 2011) because the country-by-country rankings "surprised" some countries (like Finland), while "shocking" some others (like Germany) (Grek, 2009), and "promoting" countries like the United Kingdom (UK). Martens and Niemann (2013) argued that the gap between self-perception of expected performance and the actual results might explain varying country responses to PISA.

In his overview of the impact of PISA on policy decisions, Breakspear (2012) indicated that "assessment and accountability" was the PISA policy analysis area that influenced the national policy decision most. In terms of monitoring student achievement, further national assessments were planned in many countries, including Austria, Japan, Slovak Republic, and Ireland. Some countries like Canada, Hong Kong, and Spain use PISA results as a complementary indicator to their own national results. Curriculum standards have been aligned with respect to PISA-like competencies in many countries. Educational reforms (as in Turkey, Mexico and France) were justified based on the country's PISA performance.

Comparability Issues of PISA Reading Tests

Overall, the influential role of PISA in national policy decisions seems to be based on the general acceptance that PISA tests are reliable and valid instruments because they are equivalent measures across all languages and countries and, thus, legitimately provide international comparison of student performance. However, previous research has shown that many factors such as; translation, item content/format familiarity, curricular differences, examinee motivation or test anxiety, extreme response bias, test/instrument design, sampling, calibration procedures, administration conditions, writing systems, and cultural/linguistic diversity may obscure the comparability of scores and, consequently, endanger the validity of these studies (Arffman, 2002; Bonnet, 2002; Elosua & López-Jaúregui, 2007; Grisay & Monseur, 2007; Hambleton, Merenda, & Spielberger, 2005; He & van de Vijver, 2012; Kreiner & Christensen, 2014; Mazzeo & von Davier, 2008; Oliveri & von Davier, 2011; Walker; 2007; Wetzel & Carstensen, 2013). Arffman (2010) identified six types of problems jeopardizing the equivalence of PISA reading texts. These were language specific differences in grammar, language specific differences in writing, language specific differences in meaning, differences in culture, translators' choices and strategies, and problems with editing.

Most recently, Kreiner and Christensen (2014), based on their analyses of PISA 2006 reading items, claimed that the scaling model was inadequate because of item DIF and, consequently, the ranking of countries was not viable. Some critics, consistent with concerns identified above about culture and language, have argued that PISA reading texts favor western countries to some extent, leading to a significant gap between Indo-European and non-Indo-European languages (especially for Asian, Middle Eastern, and low-GDP countries) (Grisay et al., 2007; Grisay & Monseur, 2007; Grisay, Gonzalez, & Monseur, 2009; Oliveri & von Davier, 2011). Consistent with the possibilities we have raised, it has been argued that countries with similar linguistic and cultural backgrounds were likely to exhibit equivalence in scores (Asil & Gelbal, 2012; Kankaraš & Moors, 2013). Therefore, establishing measurement invariance (MI) for PISA assessments seems problematic, though these issues may have been addressed in more recent iterations of the PISA reading literacy tests.

A further concern is the degree of invariance between different language versions within the same country and the equivalence between countries or regions using the same language. For example, Elosua and Mujika (2013) found metric invariance among the four continental Spain language versions (i.e., Spanish, Basque, Catalan, and Galician) of PISA 2009 reading literacy scale supporting score comparability across languages within one country. On the other hand, for PISA 2000 reading, the number of DIF items was found to be less within groups of countries sharing the same language (e.g., New Zealand, Ireland, USA), but much greater non-invariance has been found between different language versions used in the same country (e.g., Canada-English/French or Switzerland-German/French) (Grisay & Monseur, 2007; Grisay, Gonzalez, & Monseur, 2009).

These studies suggest that ensuring the equivalence of reading tests is often difficult to achieve. As Grisay and Monseur, (2007) pointed out; “there is always at least a basic cost in terms of loss of equivalence” (p. 82), when a test is translated. It may never be possible to achieve full equivalence across multilingual texts, meaning a high level of comparability may not be obtainable (Arffman, 2010).

Measurement Invariance (MI)

Given the broad impact of PISA and the complexity of the interrelatedness of language, culture, and education systems, the purpose of this study was to examine the equivalence of PISA 2009 reading items to evaluate the extent of score comparability between participating countries/economies. Measurement invariance (aka measurement equivalence) refers to "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" (Horn & McArdle 1992, p. 117). The lack of equivalence at the item level is referred to as differential item functioning (DIF) in the item response theory (IRT) literature (Stark, Chernyshenko, & Drasgow, 2006); while in the structural equation modeling (SEM) literature invariance is the preferred term.

Even though the need to provide "equivalence" evidence of test scores across multiple language versions of assessments is recommended by International Testing Commission (ITC, 2000) and the Standards for Educational and Psychological Testing (AERA et al., 1999) and has been investigated increasingly in the literature, it is very surprising to see just one international large scale assessment is actually reporting scale score comparability evidence (OECD, 2010d). In order to achieve international comparability of data collected in PISA surveys, it is necessary to show that all national versions of the survey instruments are equivalent. Establishing score comparability requires measurement equivalence and measurement unit equivalence which indicates the score scales on multiple versions of the tests have identical units. Scalar equivalence is required in order to compare scores from different language versions of tests in a meaningful and valid way (Ercikan & Lyons-Thomas, 2013).

MI is generally established within the framework of either IRT or Multi-Group Confirmatory Factor Analysis (MG-CFA) within SEM. In this paper we preferred to use MG-CFA because of its better performance with polytomous data (Stark, Chernyshenko, &

Drasgow, 2006). All invariance tests were based on Means and Covariance Structure (MACS) model. The recommended hierarchically structured (nested) steps (Vandenberg & Lance, 2000) of invariance tests for measurement models using multi-group applications of CFA are as follows:

1. Configural invariance: Test of equal factor structure
2. Metric (weak) invariance: Test of equal factor loadings
3. Scalar (strong) invariance: Test of equal item intercepts

Alternatively, as argued by Muthén and Asparouhov (2002) and Stark, Chernyshenko, and Drasgow (2006), after establishing configural invariance, factor loadings and intercepts (or thresholds) can be examined in tandem. These authors indicated that: (a) item probability curves are influenced by both parameters simultaneously, (b) subsequent examination increases number of comparisons which may result in higher Type I error rates, and (c) item non-invariance or non-equivalence of loadings and/or intercepts (or thresholds) is unimportant from a practical point of view. Since cross-cultural comparability requires scalar invariance, in this study we tested metric and scalar invariance at the same time; that is, loadings and thresholds were constrained (to be equal across groups) simultaneously after establishing model configural invariance. For model identifications purposes (Muthén & Muthén, 2013), we fixed the factor loading of the first item (referent item) to 1 allowing factor variances to be free across groups. Factor means were fixed at zero in all groups in the configural model; whereas, in the scalar model means were fixed at zero in the reference group and were freed in other groups.

While MG-CFA can handle multiple groups, it is more informative to conduct pairwise comparisons so as to detect the degree of similarity of each participant to a single known reference group. In terms of an international comparative analysis, any country could be arbitrarily assigned as reference point and, thus, all comparisons would be relative to the characteristics of that country's educational system, language, or socio-economic status. Distance from a common reference point, then, only indicates how similar each group is to the reference and inferences can be drawn as to how similar countries are to each other in light of their distance from the reference. In this study, it was decided to use Australia as the reference point because the Australian Council for Educational Research (ACER) coordinated and monitored the test development activities for PISA 2009. In addition, the first field testing of the booklets took place in Australian schools (OECD, 2012). However, this selection is still an arbitrary choice and the issue of whether the results would be same with a different reference (e.g., Shanghai as the highest scoring performer recently) remains to be investigated. Hence, our research question was: What type of and what magnitude of invariance or non-invariance is seen in the PISA Reading Comprehension test relative to performance by the Australian English speaking reference group and can patterns be related to language, economic, or culture differences? However, it should be noted that these analyses are conducted at the mean for each country; because there is considerable variance within each country, there may be ways in which the current analyses do not adequately capture the complexity of reading performance in each country.

Method

Instrument

Reading literacy was assessed using 131 items that are distributed across booklets. In PISA 2009, 13 different test booklets, each of which had a different subset of mathematics, science, and reading literacy items, were used in each country with a linked design. For the analyses, we selected Booklet 11, which was one of six booklets administered in all the participating partner countries. Booklet 11 contained 28 reading literacy items. Multiple choice items were scored 0 or 1; whereas, coding for the polytomous items ranged from 0 to 2. Reading processes measured with these items were Access and Retrieve (11 items), Integrate and Interpret (11 items), and Reflect and Evaluate (6 items). Reading literacy was assessed using various text formats, text types, and item formats.

Sample

Data were extracted from the PISA 2009 survey test of literacy in reading. In PISA 2009, 65 countries or economies implemented the assessment (with a further nine using the same assessment in 2010). The reading literacy tests were translated or adapted into 50 different languages. Our aim was to include the maximum number of countries so as to determine if a global picture could be derived from systematic pair-wise countries of all participating nation groups relative to the performance of the Australian reference group. After PISA 2009 assessment, some of the items (OECD, 2012, p. 196) had to be deleted at the national level due to translation, data entry, printing or layout errors, or poor item functioning. To be able to compare countries on the same scale –using same items-, countries with nationally deleted Booklet 11 reading items were excluded from the analyses. Additionally, Liechtenstein had to be removed from the data because of its small sample size (*N*=27). Thus, the effective total sample size for this study was 32,704 from 55 countries.

Data Analyses

Data were analyzed in three stages mainly using Mplus 7.1 (Muthén & Muthén, 1998-2012) software. In the first stage, Confirmatory Factor Analysis (CFA) was employed to test and confirm the unidimensional factor structure of PISA reading items. Statistical goodness of fit of the measurement model was assessed by employing multiple criteria (Cheung & Rensvold, 2002; Fan & Sivo, 2005; 2007; Hu & Bentler, 1999; Vandenberg & Lance, 2000; Wu, Li, & Zumbo, 2007). Acceptable and good model fit standards were determined by non-significant χ^2, root mean square error of approximation (RMSEA), and standardized root mean square residuals (SRMR) with values less than .08 (acceptable) or .05 (good), and the comparative fit index (CFI) and Tucker-Lewis index (TLI) with values > .90 (acceptable) or

.95 (good). We conducted separate CFA tests for the reference group and for the combined sample.

After confirming the fit of the measurement model, in the second stage, various invariance levels were assessed between Australia and each country in the dataset. MG-CFA based on Means and Covariance Structure (MACS) (Sörbom, 1974) were conducted in order to assess whether reading items functioned similarly across countries. Configural and scalar invariance of the reading items were examined to establish that valid comparisons could be made. Following the current conventions for assessing MI (Cheung & Rensvold, 2002; Vandenberg & Lance, 2000; Wu, Li, & Zumbo, 2007), the decision rule applied in this paper was ΔCFI ≤ .01, since the chi-square test is overly sensitive to large sample sizes, a necessity for this study. Additionally, ΔCFI has been shown (Chen, 2007) to be superior to ΔRMSEA and ΔSRMR, both of which are alternative indices of MI.

Although researchers have made substantial progress on assessing MI, the use of significance tests or comparing indexes with cutoffs has been criticized because such approaches dichotomize the decision rule, their sampling distribution is unknown, and, more importantly, the analyses do not indicate the magnitude of the invariance and do not address the issue of practical significance or effect size (Kirk, 2006). Therefore, in the third stage of this study, we examined the magnitude of measurement non-invariance by calculating the effect size index (d_{MACS}) for each item using dMACS computer program (freely available online) developed by Nye and Drasgow (2011). Instead of reporting the effect sizes for all 28 items and for all countries, we calculated and reported the mean d_{MACS} for each country. Interpretation of the effect size was based on Cohen's (1988) guidelines where thresholds for small, moderate, and large effects are 0.2, 0.5, and 0.8 respectively.

PISA uses a two-stage stratified sampling design where schools are sampled within countries, then students within schools. One should take complex sampling features of such data into account to compute correct standard errors and chi-square tests of model fit or MI. Therefore, Mplus "type is complex" with "weight", "stratification", and "cluster" options were used in this study for both CFA and MG-CFA analyses. Final student weight (W_FSTUWT), randomized final variance stratum (WVARSTRR) and School ID (SCHOOLID) variables from PISA database were used for this purpose.

We used robust maximum likelihood estimation method (MLR) which provides robust standard errors and adjusted $\Delta\chi^2$ when data do not follow normal distribution (Sass, Schmitt, and Marsh, 2014). Missing responses are not removed from the data but kept in the analyses using a model-based approach since MLR provides unbiased parameter estimates with missing data when they are missing at random (MAR). A graph of scalar invariance ΔCFI values against the magnitude of nonequivalence (mean d_{MACS}), relative to Australia was used to identify and interpret invariance patterns with respect to language and country.

Results

Measurement Model Data Fit

Separate CFA for reference group and combined data was performed using the MLR estimation method. As can be seen from Table 1, the one factor model of PISA reading test consisting of 28 items, provided acceptable (i.e., CFI/TLI values > .90) to good (RMSEA and SRMR values were < .05) fit for the Australian sample and for the combined sample.

<<<Insert Table 1 about here>>>

After confirming the factor structure of the reading performance, parameter estimates for the measurement model were obtained with MLR (Table 2).

<<<Insert Table 2 about here>>>

Standardized factor loadings, all of which were statistically significant, ranged from .219 to .654 for the reference group, from .261 to .627 for the combined sample, suggesting equivalence at the metric level was plausible. For both Australian and the combined sample, the internal consistency reliability estimate of the measurement model was α=.89, which is greater than the commonly accepted threshold value of α=.70. After establishing adequate model data fit, we conducted multi-group confirmatory factor analyses (MG-CFA) within the framework of MACS.

Measurement Invariance (MI) Analyses

Measurement invariance analyses results for each comparison of Australia and every other country are summarized in Table 3. The results showed that the configural model fit the data reasonably well (range of χ^2/df: 1.713 - 2.500; range of RMSEA: .024 - .045, range of CFI: .848- .924, and range of SRMR of .024 - .052) for all 54 comparisons. This indicated that students from different countries used the same conceptual framework to answer the reading items and that further investigation of MI was warranted. Scalar invariance, on the other hand, showed that intercepts (with constrained factor loadings) were not invariant (ΔCFI>.01) for almost all countries compared to Australia. Only three countries (i.e., New Zealand, Canada, and USA) in our data had scalar invariance to Australia.

<<<Insert Table 3 about here>>>

Effect Size Analyses

To address the practical importance of the observed nonequivalence of each country to Australia, the magnitude of invariance for each item (d_{MACS}) was calculated in dMACS with the unstandardized parameters from the configural invariance models. The effect size index (d_{MACS}) takes both loadings and intercepts into account simultaneously. For the ease of reporting, mean d_{MACS} values were calculated for each comparison and are also provided in Table 3. The effect size measures ranged from d_{MACS}=.041 (insubstantial effect) to d_{MACS}=.928 (large effect).

A graph of the scalar invariance ΔCFI values against the mean d_{MACS} effect size measures of nonequivalence relative to Australia was generated to assist in identifying patterns in the invariance. Figure 1 shows the order of countries relative to Australia based on the difference in CFI for scalar invariance and country means of d_{MACS} index.

<<<Insert Figure 1 about here>>>

Despite statistically significant differences at the scalar level, the effect size analyses indicated that 43% of the effect sizes between Australia and other countries were insubstantial, 26% were small, and only 31% were moderate to large. In total, 23 countries, including three which had scalar invariance, had trivial effect size differences (mean d_{MACS} <.20) to Australia. These included the four wealthy English-speaking countries, 12 countries of Western Europe (plus Estonia), and six high-performing East Asian jurisdictions (i.e., Japan, Taipei, Korea, Shanghai, Hong Kong, and Macau). These countries are the predominantly wealthy nations participating in the survey which invest considerable resources in education or have high cultural emphasis on educational performance. There are clearly no patterns here of impact to do with language family, writing script, or culture. Likewise, the 16 countries with effect sizes in the moderate to large range (i.e, mean d_{MACS} >.50) had a variety of scripts (albeit all syllabic), locations, and cultures; instead, this group of South American, Eastern European, Asian, and Middle Eastern countries seem to have relatively lower levels of investment in education.

To test the effect of socio-economic status, we derived the PISA index of economic, social and cultural status (ESCS) from OECD (2010c) report. The ESCS index "captures a range of aspects of a student's family and home background that combines information on parents' education and occupations and home possessions" (OECD, 2010c, p.29). The relationship between ESCS and ΔCFI and d_{MACS} for the 47 countries for which it was available was investigated (OECD, 2010c). There was a moderate but negative relationship between ESCS and ΔCFI ($r = -.61$, $p < .05$), ESCS and d_{MACS} ($r = -.54$, $p < .05$), indicating that lower levels of ESCS tended to be associated with less equivalence to Australia and much greater effect sizes in the difference.

Figure 1 clarifies that language, script, and culture are not strong factors in explaining invariance; differences in socio-economic resources seem important both within and across language groups. For example, Trinidad-Tobago uses English, but is not a high-wealth society, and had a moderately-large effect size relative to Australia (dMACS=.55). Similarly, Portugal ($8000USD per pupil expenditure), the richer country, using the same language as Brazil ($3000USD per pupil), was considerably closer to Australian parameters (dMACS=.16 vs. dMACS=.53, respectively).

In terms of invariance to the Australian model, language similarity seems to play a small role in these results. We observed that Indo-European languages were relatively located in the bottom half of the graph; whereas, non-Indo-European languages tended to be in the upper half, which is consistent with our hypothesis about the similarity of languages influencing reading achievement. Nonetheless, this is a much weaker contributor to observed differences than the impact of socio-economic resources. Likewise, it seems that the type of writing script used in different languages might be a small contributor to non-invariance. Most of the languages in the bottom half of Figure 1 use a Roman or Latin alphabet; whereas, we see Cyrillic, Arabic, and Chinese scripts mostly in the upper half of scalar invariance.

This provides some support for the hypothesis that changes in the nature of reading comprehension arise in response to differences in reading non-syllabic or phonemic scripts. Hence, these results suggest that, once reading for comprehension is mastered, impact on models of reading comprehension are minimal, unless exacerbated by significant differences in socioeconomic and cultural resources.

Furthermore, British Commonwealth countries that emphasize a child-centered pedagogical approach and which were relatively wealthy were invariant to Australia. In contrast, more traditional societies, probably emphasizing more didactic teaching, seemed to group at the top of the graph, relatively variant to Australia, but only in terms of scalar invariance, rather than effect size. This suggests that, insofar as reading literacy in an achievement test context is concerned, approaches to teaching are less consequential than commonly thought. Efforts to change pedagogical practices in such contexts to more child-centered approaches may not make any substantial difference to performance on PISA.

Results indicate that complex factors to do with educational practice and socio-economic resourcing of education, rather than language or writing per se do interfere with the MI of the PISA reading comprehension results, though for the most part these are not practically significant, with impact seen most strongly among the poorer economies.

Discussion and Conclusion

This study advances our understanding of the nature of invariance in the PISA reading literacy tests by comparing performance in 55 countries and replicates many previous studies reporting lack of MI in PISA tests (e.g., Arffman, 2010; He & van de Vijver, 2012; Kreiner & Christensen, 2014; Oliveri & von Davier, 2011; Wetzel & Carstensen, 2013). We examined the invariance of parameters in a factor analysis model used to represent the performance of 55 countries on the 2009 PISA reading literacy test. A one factor model of reading literacy scale reported by PISA was good fitting and configurally invariant across countries. MG-CFA with nested invariance testing showed that only three countries had scalar invariance to the Australian model. Inspection of the degree of invariance to Australia showed that the invariant samples were from countries which were similar to Australia predominantly in terms of socio-economic resources. Using a non-European language, using a non-Roman script, and having a more transmission-oriented pedagogical approach seemed to make a small contribution to scalar non-invariance.

However, the magnitude analysis of the non-invariance revealed that 69% of the comparisons exhibited trivial to small effects. Given the differences in participating countries' languages, cultures, educational systems and economies, small differences may be acceptable for PISA assessments. However, there still remain 31% of economies with moderate and large effects relative to the performance of Australian students. Given the correlation with d_{MACS} effect size, it seems that invariance from Australia on a meaningful scale is driven predominantly by ESCS, with a complex interaction with, albeit small contribution from educational practice, cultural factors, and language/writing itself. The more an education system and economy is similar to Australia, the more likely its students will respond to the PISA reading literacy tests in a similar and comparable fashion.

From these results, we conclude that the PISA reading scale scores can legitimately be compared to Australia when ESCS is reasonably similar; but large differences in ESCS are reflected in substantial differences in how students respond to PISA reading comprehension items. Hence, it suggests that, in reflecting upon the legitimacy of the large policy impact of PISA upon countries, rather than abandoning attempts to compare performance, PISA might better address the lack of invariance by reporting results more cautiously. At least a two-tier system might be considered to better account for the large impact of ESCS (i.e., rankings for high vs. low ESCS economies). On the other hand, among the developed nations there were small effects attributable to language group, with very similar differences to Australia seen in jurisdictions using the same language. This suggests that reporting PISA results by language group (e.g., English, Spanish, French, German, etc.) and making comparisons only within language groups that demonstrate invariance would more likely result in legitimate evaluations of performance and less drastic policy impacts upon already stressed educational systems. Certainly, the Anglo-Commonwealth countries in this study were invariant to Australia and ranking their relative performances seems defensible. The closeness of Canada and the United States to each other and their small effect size difference in invariance relative to Australia does suggest that a different reference point may make the responses of all the English, developed nations invariant. In other words, if Canada, with its federal system and tendency to use high-stakes testing to judge schooling quality, had been used as a reference, it may be that all the English speaking, high ESCS countries would have been invariant. This would mean comparisons of mean score amongst those countries would become completely legitimate.

Alternately, grouping countries into clusters of ‘countries-like-me’, assuming they use the same language and/or share the same educational culture, might be a better way to identify invariance of responding to test items and to report results. Restricting reports to those nations would likely be defensible and informative. For example, it may be conventional to argue that East Asian societies which use Mandarin (i.e., Singapore, China, Macao, and Taiwan) and having strong dependence on testing and public examinations and shared cultural approaches to schooling and testing should be compared, independent of other East Asian societies which have different writing scripts and languages, despite having similar cultural histories and forces (i.e., Japan, Korea, and Hong Kong). However, this study has suggested that grouping and reporting performances among East Asian societies may be defensible, since the range of d_{MACS} relative to Australia for all seven economies was just .136 to .199. Likewise, once could imagine Nordic countries (i.e., Finland, Sweden, Norway, Iceland, and Denmark) or continental Western European countries choosing to allow inter-country comparisons because of their similarities in ESCS, despite differences in language. Nonetheless, separate studies that demonstrate that such natural geographic and cultural groupings were defensible requires conducting parallel analyses using one or more of the contributing nations as a reference point.

As suggested by the many previous studies of invariance within country or language or region, this study found that the 2009 PISA reading literacy tests was not an invariant test between Australia and 54 other economies. While every effort is being made to maximize the psychometric basis of comparison, the current approach adopted by PISA does not achieve equivalence; albeit for the more developed jurisdictions the scale of non-invariance was trivial to small. We suspect that a globally invariant test of reading literacy may be an impossibility given the many different factors that impinge upon reading instruction and reading performance. We suggest, instead, that PISA concentrate upon developing reports and rankings which are justifiably comparable and this study points us to a possible solution that should be considered.

Limitations and Suggestions

A number of limitations to this study point to further investigations. Only one booklet was used in the current analyses. It was not possible with the confirmatory factor analytic approach to use all the booklets simultaneously because of the planned missing data structure of the booklet design. Hence, results from Booklet 11 may not generalize to all PISA reading literacy items or tasks. Further, in some multilingual countries (e.g., Canada, Switzerland, and Belgium) multiple language versions of Booklet 11 were used. Despite previous research (Grisay & Monseur, 2007; Grisay, Gonzalez, & Monseur, 2009) that non-invariance exists between different language versions within the same country, this aspect couldn't be tested and isolated because of sampling restrictions. Hence, further studies could examine the type and scale of invariance within nations across languages.

It is also well established in the literature that the selection of a referent item from within the test booklet may play an important role in MI analyses and effect size calculations. It is very likely that the referent item chosen in this study may not be fully equivalent or unbiased across all comparisons for reasons inherent to the specific nature of countries themselves. Therefore, further analysis of ΔCFI and d_{MACS} need to be conducted to establish their characteristics when the sample size is small and when different items are chosen as the anchor.

Meaningful and valid cross-country or cross-lingual comparisons in international surveys require invariant measurement parameters which most of the time may be an unrealistic assumption. Further studies may employ recently developed promising approaches like alignment method (Muthen & Asparouhov, 2013a) or Bayesian structural equation modeling measurement invariance analysis (Muthen & Asparouhov, 2013b) which assume only approximate measurement equivalence and, yet, can optimally estimate parameters on latent variables.

Nonetheless, this study has shown that at least one PISA OECD reading comprehension test generally fails the assumptions of scalar equivalence relative to the Australian baseline, although for most countries this difference was trivial to small. However, for about a third of countries, characterized generally by lower ESCS, this sample test reflects a substantially different model of reading literacy, making international comparisons questionable. The paper has suggested some alternative analyses and even reporting solutions to mitigate invalid interpretations for such contexts.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Arffman, I. (2002). *In search of equivalence: Translation problems in international literacy studies* (Unpublished master's thesis). University of Jyväskylä, Jyväskylä, Finland.

Arffman, I. (2010). Equivalence of translations in international reading literacy studies. *Scandinavian Journal of Educational Research, 54*(1), 37-59.

Asil, M. & Gelbal, S. (2012). PISA öğrenci anketinin kültürler arası eşdeğerliği. *Egitim ve Bilim*, *37*(166), 236-249.

Baird, J., Isaacs, T., Johnson, S., Stobart, G., Yu, G., Sprague, T., & Daugherty, R. (2011). *Policy effects of PISA*. Oxford University Centre for Educational Assessment.

Bonnet, G. (2002). Reflections in a critical eye: on the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice*, *9*(3), 387-399.

Boroditsky, L. (2001). Does language shape thought? English and Mandarin speakers' conceptions of time. *Cognitive Psychology, 43*(1), 1-22.

Boroditsky, L. (2011). How language shapes thought. *Scientific American, 304*(2), 62-65.

Boroditsky, L., Fuhrman, O., & McCormick, K. (2010). Do English and Mandarin speakers think differently about time? *Cognition*, *118*(1), 123-129.

Boroditsky, L. & Gaby, A. (2010). Remembrances of times east: Absolute spatial representations of time in an Australian Aboriginal community. *Psychological Science, 21*(11), 1635-1639.

Breakspear, S. (2012). *The policy impact of PISA*. OECD Education Working Paper 71. Paris: OECD.

Brown, G. T. L., & Wang, Z. (2013). Illustrating assessment: how Hong Kong university students conceive of the purposes of assessment. *Studies in Higher Education, 38*(7), 1037-1057. doi: 10.1080/03075079.2011.616955

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of 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*(2), 233-255.

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

Elosua, P., & López-Jaúregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing, 7*(1), 39-52.

Elosua O. P., & Mujika L. J. (2013). Invariance levels across language versions of the PISA 2009 reading comprehension tests in Spain. *Psicothema*, *25*(3), 390-395.

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting Tests for Use in Other Languages and Cultures. In K. Geisinger (Ed), *APA Handbook Testing and Assessment in Psychology* (vol. 3, pp. 545-569). American Psychological Association: Washington.

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12*(3), 343-367.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509-529.

Fausey, C. & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic bulletin & review*, *18*(1), 150-157.

Fromkin, V., & Rodman, R. (1978). *An introduction to language* (2nd ed.). New York: Holt, Rinehart and Winston.

Grek, S. (2009). Governing by numbers: The PISA ‘effect’in Europe. *Journal of Education Policy, 24*(1), 23-37.

Grisay, A., de Jong, J. H., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249-266.

Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLS and PISA reading assessments. *IERI monograph series: Issues and methodologies in large-scale assessments, 2*, 63-84.

Grisay, A. & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation 33*(1), 69-86.

Hambleton, R.K., Merenda, P., & Spielberger, C. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence S. Erlbaum Publishers.

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*(2). http://dx.doi.org/10.9707/2307-0919.1111

Hofstede, G. (2007). A European in Asia. *Asian Journal of Social Psychology, 10*(1), 16-21. doi: 10.1111/j.1467-839X.2006.00206.x

Hofstede, G., & Bond, M. H. (1988). The Confucius connection: From cultural roots to economic growth. *Organizational Dynamics, 16*(4), 5-21. doi: http://dx.doi.org/10.1016/0090-2616(88)90009-5

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

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

ITC (2000). *ITC Guidelines for adaptation*. Retrieved from http://www.intestcom.org/test_adaptation.htm

Kankaraš, M., & Moors, G. (2013). Analysis of Cross-Cultural Comparability of PISA 2009 Scores. *Journal of Cross-Cultural Psychology, 45*(3), 381-399.

Kirk, R. E. (2006). Effect magnitude: A different focus. *Journal of Statistical Planning and Inference, 137,* 1634-1646.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. a new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210-231.

Martens, K., & Niemann, D. (2013). When Do Numbers Count? The Differential Impact of the PISA Rating and Ranking on Education Policy in Germany and the US. *German Politics, 22*(3), 314-332.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results.* OECD Education Working Papers EDU/PISA/GB, 28.

McWhorter, J. (2003). *The power of Babel*. London: Arrow Books.

Muthén, L.K. & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus. Retrieved from http://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, L.K. & Asparouhov, T. (2013a). New methods for the study of measurement invariance with many groups. Retrieved from http://www.statmodel.com/download/PolAn.pdf

Muthén, L.K. & Asparouhov, T. (2013b). BSEM measurement invariance analysis. Retrieved from http://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, L.K. & Muthén, B.O. (1998-2012). *Mplus user's guide 7*. Los Angeles, CA: Muthén & Muthén.

Muthén L. K. & Muthén B. O. (2013). *Version 7.1 Mplus language addendum*. Available online at: http://www.statmodel.com/ugexcerpts.shtml.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analysis of measurement equivalence: Understanding the practical importance of difference between groups. *Journal of Applied Psychology, 96,* 966-980. doi: 10.1037/a0022955

OECD (2010a). *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*, OECD, Paris, France.

OECD (2010b). *PISA 2009 results: Learning trends: Changes in student performance since 2000 (Volume V),* OECD, Paris, France.

OECD (2010c). *PISA 2009 Results: Overcoming social background – equity in learning opportunities and outcomes (Volume II),* OECD, Paris, France.

OECD (2010d). *TALIS technical report*. Paris, France.

OECD (2012). OECD (2012), *PISA 2009 Technical Report*, PISA, OECD Publishing. http://dx.doi.org/10.1787/9789264167872-en

OECD (2013). *Education at a Glance 2013: OECD Indicators* doi: 10.1787/eag-2013-en

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Journal of Psychological Test and Assessment Modeling, 53*(3), 315-333.

Perfetti, C. A., & Harris, L. N. (2013). Universal reading processes are modulated by language and writing system. *Language Learning and Development, 9*(4), 296-316. doi: 10.1080/15475441.2013.813828

Peterson, E. R., Brown, G. T. L., & Hamilton, R. J. (2013). Cultural differences in tertiary students' conceptions of learning as a duty and student achievement. *The International*

Journal of Quantitative Research in Education, 1(2), 167-181. doi: 10.1504/IJQRE.2013.056462

Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating model fit with ordered categorical data within a measurement invariance framework: A comparison of estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167-180.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210-222.

Shuell, T. J. (1996). Teaching and learning in a classroom context. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 726-764). New York: Macmillan.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27*(2), 229-239.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Stobart, G. (2006). The validity of formative assessment. In J. Gardner (Ed.), *Assessment and learning* (pp. 133-146). London: Sage.

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*, 4-69.

Walker, M. (2007). Ameliorating culturally based extreme response tendencies to attitude items. *Journal of Applied Measurement, 8*(3), 267-278.

Wetzel, E. & Carstensen, C. H. (2013). Linking PISA 2000 and PISA 2009: Implications of instrument design on measurement invariance. *Psychological Test and Assessment Modeling, 55*(2), 181-206.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation, 12*(3), Retrieved from: http://pareonline.net/getvn.asp?v=12&n=13.

Table 1

Confirmatory Factor Analysis Results for the Reference Group and Combined Sample

Group	N	χ^2	df	SCF	RMSEA (90% CI)	CFI	TLI	SRMR
Australia	1112	760.570*	350	6.424	.033 (.029, .036)	.911	.904	.041
Combined sample	32704	2609.444*	350	2.305	.014 (.014, .015)	.932	.926	.028

Note. RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index; TLI, Tucker-Lewis Index; SRMR, Standardized Root Mean Squared Residual; SCF, Scaling Correction Factor for MLR.

* *p*<.05.

Table 2

Standardized Factor Loadings and Reliabilities for the Scale

Item Code	Australia (λ_i)	Combined Sample (λ_i)
R055Q01	0.464	0.473
R055Q02	0.535	0.479
R055Q03	0.644	0.579
R055Q05	0.654	0.627
R104Q01	0.516	0.567
R104Q02	0.250	0.306
R104Q05	0.439	0.460
R111Q01	0.564	0.524
R111Q02B	0.527	0.481
R111Q06B	0.594	0.539
R227Q01	0.457	0.428
R227Q02T	0.555	0.544
R227Q03	0.601	0.536
R227Q06	0.616	0.601
R432Q01	0.580	0.542
R432Q05	0.602	0.585
R432Q06T	0.360	0.356
R446Q03	0.386	0.369
R446Q06	0.391	0.464
R456Q01	0.219	0.261

R456Q02	0.381	0.440
R456Q06	0.460	0.456
R460Q01	0.535	0.533
R460Q05	0.440	0.515
R460Q06	0.529	0.482
R466Q02	0.471	0.537
R466Q03T	0.265	0.283
R466Q06	0.571	0.546
Reliability (α)	.885	.894

Table 3

Measurement Invariance of the Scale across Countries

	Configural Fit Statistics								Scalar Difference		Effect Size
Australia vs. Country	*N*	χ^2	*p*	SCF	RMSEA	RMSEA (90% CI)	CFI	SRMR	ΔCFI	ΔRMSEA	Average d_{MACS}
Austria	490	1308.637	.000	2.050	0.033	.030, .036	0.906	0.043	-0.032	0.004	0.176
Belgium	618	1273.792	.000	1.994	0.031	.028, .033	0.921	0.040	-0.031	0.004	0.128
Brazil	1536	1349.151	.000	2.656	0.026	.024, .029	0.916	0.040	-0.072	0.009	0.528
Bulgaria	337	1396.053	.000	1.807	0.037	.034, .040	0.899	0.044	-0.059	0.008	0.500
Canada	1758	1658.180	.000	2.834	0.031	.029, .033	0.890	0.043	**-0.004**	-0.001	0.041
Chile	417	1342.019	.000	1.827	0.035	.032, .037	0.906	0.042	-0.042	0.005	0.275
Chinese Taipei	450	1322.688	.000	2.147	0.034	.031, .037	0.907	0.042	-0.055	0.007	0.172
Colombia	627	1198.913	.000	2.276	0.029	.026, .031	0.913	0.043	-0.087	0.010	0.510
Costa Rica	350	1324.326	.000	1.962	0.035	.032, .038	0.904	0.044	-0.039	0.005	0.394
Croatia	397	1361.417	.000	1.832	0.035	.033, .038	0.907	0.042	-0.067	0.010	0.279
Czech Republic	455	1321.384	.000	2.122	0.034	.031, .036	0.902	0.043	-0.030	0.003	0.218

	Configural Fit Statistics								Scalar Difference		Effect Size
Australia vs. Country	*N*	χ^2	*p*	SCF	RMSEA	RMSEA (90% CI)	CFI	SRMR	ΔCFI	ΔRMSEA	Average d_{MACS}
Denmark	447	1253.295	.000	2.119	0.032	.029, .035	0.911	0.042	-0.048	0.006	0.153
Estonia	369	1284.167	.000	2.303	0.034	.031, .036	0.905	0.044	-0.055	0.007	0.136
France	339	1367.799	.000	1.829	0.036	.033, .039	0.905	0.043	-0.024	0.003	0.198
Georgia	364	1348.277	.000	1.766	0.035	.033, .038	0.905	0.043	-0.081	0.011	0.639
Germany	366	1300.882	.000	2.081	0.034	.031, .037	0.911	0.043	-0.047	0.007	0.193
Greece	380	1347.779	.000	1.878	0.035	.032, .038	0.906	0.043	-0.021	0.003	0.356
Hong Kong-China	367	1259.864	.000	2.196	0.033	.030, .036	0.908	0.043	-0.068	0.009	0.171
Iceland	286	1310.157	.000	2.207	0.035	.032, .038	0.908	0.043	-0.024	0.003	0.110
Indonesia	399	1388.689	.000	2.925	0.036	.033, .039	0.892	0.044	-0.085	0.010	0.528
Italy	2378	1598.914	.000	2.442	0.027	.025, .029	0.918	0.036	-0.041	0.005	0.155
Japan	472	1571.515	.000	1.992	0.040	.037, .042	0.887	0.047	-0.053	0.006	0.152
Jordan	496	1397.599	.000	1.984	0.035	.033, .038	0.894	0.044	-0.104	0.013	0.619
Korea	379	1460.383	.000	2.893	0.038	.035, .041	0.876	0.047	-0.056	0.006	0.145

Australia vs. Country	Configural Fit Statistics								Scalar Difference		Effect Size
	N	χ^2	p	SCF	RMSEA	RMSEA (90% CI)	CFI	SRMR	ΔCFI	ΔRMSEA	Average d_{MACS}
Kyrgyzstan	386	1391.032	.000	2.192	0.036	.034, .039	0.896	0.045	-0.142	0.018	0.928
Latvia	350	1214.552	.000	2.104	0.032	.029, .035	0.909	0.043	-0.024	0.002	0.248
Luxembourg	360	1350.657	.000	1.778	0.036	.033, .038	0.908	0.043	-0.036	0.004	0.187
Macao-China	455	1299.163	.000	1.974	0.033	.030, .036	0.908	0.041	-0.073	0.010	0.136
Malaysia	390	1408.877	.000	1.789	0.037	.034, .040	0.897	0.044	-0.083	0.011	0.544
Malta	267	1445.444	.000	1.659	0.039	.036, .042	0.900	0.045	-0.040	0.006	0.544
Mauritius	349	1363.209	.000	1.806	0.036	.033, .039	0.901	0.045	-0.069	0.009	0.587
Mexico	2948	1523.523	.000	2.582	0.024	.022, .026	0.924	0.024	-0.100	0.011	0.432
Miranda-Venezuela	232	1281.125	.000	2.030	0.035	.032, .038	0.905	0.045	-0.043	0.006	0.593
Netherlands	347	1226.203	.000	2.359	0.032	.029, .035	0.911	0.043	-0.050	0.007	0.171
New Zealand	355	1427.454	.000	2.193	0.038	.035, .040	0.896	0.046	**-0.003**	-0.001	0.064
Norway	361	1382.943	.000	1.865	0.036	.034, .039	0.902	0.044	-0.043	0.006	0.218
Panama	302	1223.647	.000	2.294	0.033	.030, .036	0.898	0.048	-0.065	0.007	0.617

Australia vs. Country	Configural Fit Statistics								Scalar Difference		Effect Size
	N	χ^2	*p*	SCF	RMSEA	RMSEA (90% CI)	CFI	SRMR	ΔCFI	ΔRMSEA	Average d_{MACS}
Peru	465	1342.216	.000	1.849	0.034	.031, .037	0.911	0.043	-0.095	0.013	0.733
Portugal	500	1309.308	.000	1.903	0.033	.030, .036	0.915	0.042	-0.039	0.005	0.157
Qatar	702	1565.876	.000	1.724	0.037	.035, .039	0.909	0.041	-0.079	0.012	0.823
Republic of Moldova	394	1750.332	.000	1.787	0.045	.042, .047	0.848	0.052	-0.079	0.008	0.626
Romania	372	1372.171	.000	1.819	0.036	.033, .039	0.901	0.044	-0.061	0.008	0.527
Russian Federation	412	1387.005	.000	1.920	0.036	.033, .039	0.896	0.044	-0.077	0.010	0.343
Shanghai-China	398	1226.946	.000	2.742	0.032	.029, .035	0.914	0.042	-0.065	0.008	0.199
Slovak Republic	358	1334.889	.000	1.896	0.035	.032, .038	0.906	0.043	-0.031	0.004	0.206
Slovenia	457	1336.168	.000	1.993	0.034	.031, .037	0.906	0.043	-0.037	0.005	0.150
Spain	2002	1429.714	.000	2.947	0.026	.024, .028	0.912	0.039	-0.049	0.005	0.219
Switzerland	881	1496.007	.000	2.477	0.034	.031, .036	0.880	0.046	-0.028	0.002	0.167
Thailand	473	1299.240	.000	1.850	0.033	.030, .036	0.909	0.042	-0.097	0.012	0.494
Trinidad and Tobago	379	1450.271	.000	1.879	0.038	.035, .041	0.896	0.046	-0.039	0.005	0.546

	Configural Fit Statistics								Scalar Difference		Effect Size
Australia vs. Country	*N*	χ^2	*p*	SCF	RMSEA	RMSEA (90% CI)	CFI	SRMR	ΔCFI	ΔRMSEA	Average d_{MACS}
Turkey	389	1347.657	.000	1.766	0.035	.032, .038	0.904	0.043	-0.038	0.005	0.304
United Kingdom	957	1270.279	.000	2.667	0.028	.026, .031	0.916	0.042	-0.014	0.001	0.123
United States	407	1258.218	.000	2.183	0.032	.030, .035	0.918	0.042	**-0.007**	0.000	0.137
Uruguay	467	1423.530	.000	1.773	0.036	.034, .039	0.902	0.044	-0.027	0.003	0.453

Note. RMSEA, Root Mean Square Error of Approximation; CFI, Comparative Fit Index; SRMR, Standardized Root Mean Squared Residual; SCF, Scaling Correction Factor for MLR, Degrees of freedom are 700 for all models, ΔCFI ≤ .01 values were highlighted in bold.

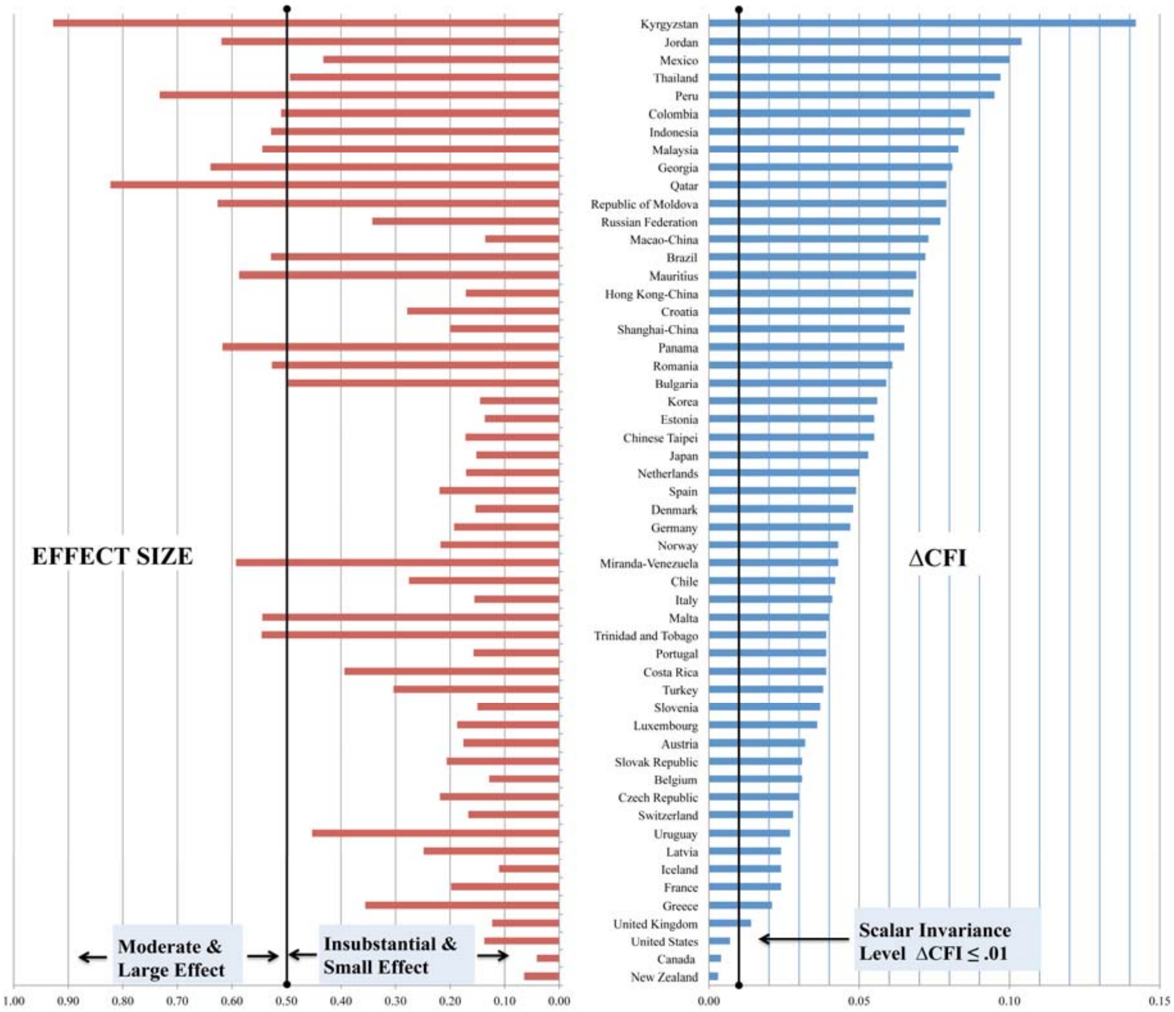

Figure 1. Countries by scalar invariance difference from Australia and mean d_{MACS} index.