Measurement Instruments for the Social Sciences

MEETING REPORT Open Access

A meeting report: cross-cultural comparability of questionnaire measures in large-scale international surveys

Francesco Avvisati, Noémie Le Donné and Marco Paccagnella*

Abstract

The value of cross-country comparisons is at the heart of large-scale international surveys. Yet the validity of such comparisons is often challenged, particularly in the case of latent traits whose estimates are based on self-reported answers to a small number of questionnaire items. Many believe self-reports to be unreliable and not comparable, and indeed, formal statistical procedures very often reject the assumption that the questions are understood and answered in the same way in different countries (measurement invariance). A methodological conference on the comparability of questionnaire scales was hosted by the OECD on 8 and 9 November 2018. This meeting report summarises the discussions held at the conference about measurement invariance testing and instrument design. The report first provides a brief introduction to the measurement models and the accompanying invariance analyses typically used in the industry of large-scale international surveys and points to the main limitations of these current standard approaches. It then presents classical and novel ways to deal with imperfect comparability of measurements when scaling and reporting on continuous traits and on categorical latent variables. It finally discusses the extent to which item design can improve the cross-country comparability of the measured constructs (e.g. by adopting innovative item formats such as anchoring vignettes and situational judgement test items). It concludes with some general considerations for survey design and reporting on invariance analyses and survey results.

Keywords: Measurement invariance, International large-scale assessments, IRT, Factor analysis, Latent class analysis, Situational judgement test, Anchoring vignettes

Introduction

The value of cross-country comparisons is at the heart of large-scale international surveys, including those piloted by the Organisation for Economic Co-operation and Development (OECD), such as the Programme for International Student Assessment (PISA), the Programme for the International Assessment of Adult Competencies (PIAAC), and the Teaching and Learning International Survey (TALIS). When surveys go beyond measuring objective attributes (e.g. age or household size) and behaviours (e.g. unemployment or job-seeking behaviours) and aim to assess subjective attitudes (e.g. attitudes towards migrants or subjective well-being), or psychological traits such as perseverance, new challenges for the validity and comparability of survey results emerge, and old issues acquire renewed salience. Reflective latent constructs measured through self-reports, for example, are particularly affected by subtle linguistic differences in the translated questionnaires and by broader cultural differences. These may introduce variation in participants' understanding of survey questions and therefore in the relationship between their responses and the target latent construct. Similarly, when confronted with Likert items, with generic frequency scales ("often", "sometimes", "never or almost never"), or with subjective rating scales ("on a scale from 1 to 10"), cultural norms may mediate the response process of participants. As a result, international surveys may fall short of their objective to perform comparisons across countries.

These issues of cross-cultural comparability were recently the focus of a methodological conference hosted at the OECD headquarters: How can different levels of

* Correspondence: marco.paccagnella@oecd.org
OECD, Directorate for Education and Skills 2, Rue André Pascal,, 75775 Paris Cedex 16, France

comparability be defined? How can they be identified in the data? How should violations of comparability be addressed when analysing and reporting these data, to prevent misuse of the data in policy discussions? How can instruments be designed in order to maximise comparability?

The conference brought together leading experts in questionnaire design and in the statistical modelling of survey responses with representatives from the industry involved in the development of questionnaires, data products, and reports. The objective was to identify areas where current practices for designing and analysing questionnaires in cross-national large-scale surveys can improve, while keeping in mind the practical constraints, the timelines, and the reporting goals of such surveys.

The conference tackled two main topics: first, innovative statistical methods to deal with imperfect comparability of measurements, distinguishing between the case of continuous and categorical latent traits, and second, innovative item formats (and more general design principles) that could be followed to achieve higher levels of comparability. The conference did not discuss (due to time limitations) other issues related to survey design and administration that have an important bearing on the comparability of survey results, such as consistency in sampling design or translation.

Before reporting on the discussions held at the conference, the next section provides a formal definition of measurement invariance, describes the statistical methods commonly used to assess invariance, and points to the main limitations of these "standard" approaches.

The subsequent sections report on the various presentations and discussions held at the conference, while the final section draws some conclusions on lessons learnt. A detailed conference agenda is provided in Additional file 1: Annex A.

Measurement invariance in large-scale international surveys

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

Measurement invariance can be defined as "a property of a measurement instrument (in the case of survey research, a questionnaire), implying that the instrument measures the same concept in the same way across various subgroups" (Davidov et al., 2014, p. 58). In more technical terms, this implies that the measurement model meets a conditional independence property with respect to a set of subpopulations within the parent population (e.g. countries, gender, time) (Horn & Mcardle, 1992; Mellenbergh, 1989; Meredith, 1993).

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

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

A standard of the past

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

et al., 2014; van de Schoot, Lugtig, & Hox, 2012).[1] Model-fit indices are then compared across these nested models, and conclusions are drawn about whether the data conform to the stronger "scalar invariance" hypothesis, or to the weaker "metric" or "configural" invariance hypotheses.

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

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

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

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

Excitement around new developments

In recent years, many alternative paradigms in measurement equivalence research have emerged. The conference was meant to be a forum to introduce some of the main new developments and to discuss the extent to which these could lead to the establishment of new standards in international large-scale surveys and support robust conclusions about cross-country differences.

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

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

Partial invariance

Model-building approaches are very common in the IRT framework and have often been used by MGCFA practitioners in response to the failure to establish full scalar invariance. Starting from a fully invariant (scalar invariant) model, these approaches estimate item-level fit indices for every group, identify the items for which certain groups exhibit high level of misfit (usually referred to as differential item functioning, or DIF, in IRT), and then deal with misfit by sequentially releasing constraints, until adequate fit is reached. This results in the so-called partial invariance models, whereby the conditional independence holds for some measurements (often referred to as "anchor items"),

but not all (Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998). This approach is currently in use in the PISA assessment, both in the scaling of the cognitive component (von Davier, Yamamoto, Shin, Chen, Khorramdel, Weeks, et al. &, 2018) and in the analysis of questionnaire scales (Buchholz & Hartig, 2017).

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

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

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

Alignment optimisation

In recent years, an alternative response to the failure to establish full scalar invariance in MGCFA has gained popularity, the so-called alignment optimisation approach (Davidov & Meuleman, 2019). This approach tolerates small differences, even if there are many of them. The popularity of the approach is due to its simplicity and to its availability in the popular software package Mplus (Muthén & Muthén, 1998-2017). It requires only two steps: (1) estimation of a model with group-specific parameters ("configural model") and (2) minimisation of a loss function which depends on differences between parameters across groups, leading to a "rotated" solution which forces the group means and variances from the configural model on a same scale. The procedure is similar to factor rotation in exploratory factor analysis (EFA) (Asparouhov & Muthén, 2014) and can equally be applied in the IRT context (Muthén & Asparouhov, 2014). Matthias von Davier, in particular, also highlighted how the alignment optimisation method is very similar to the simultaneous test-linking approach proposed by Haberman (2009).

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

Bayesian approximate invariance methods

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

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

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

Rens van de Schoot suggested that because of the dependence on priors, practitioners should conduct a sensitivity analysis before drawing substantive conclusions, i.e. estimate models with different priors and verify the robustness of the resulting claims (Lek & van de Schoot, 2019). In general, there was no consensus on how to rank models based on different priors (and thus, select the "best" priors and models): Jean-Paul Fox highlighted that criteria such as posterior predictive *p* values (PPP) or deviance information criteria (DIC) should not be used to compare models with the same number of parameters. On the other hand, using the same priors for all parameters may be just as unrealistic as assuming that there is no variation in measurement parameters, but tailored priors may invite an abuse of "researcher degrees of freedom" (Simmons, Nelson, & Simonsohn, 2011), especially if they influence the conclusions strongly.

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

In response to some of these shortcomings, Jean-Paul Fox presented an alternative approach to assess whether the data support full invariance or only approximate invariance of measurements, which he illustrated in the IRT case (Fox, 2019). The approach, which was recently presented in Fox, Mulder, and Sinharay (2017), is based on the intuition that the marginal model obtained by integrating out the random parameters from a one-parameter IRT model is simply a fixed-effect model with a particular structure for the covariance of residuals. Therefore, it is possible to conduct an analysis of residuals from the simpler model to identify (using Bayes Factor tests) whether a complex covariance structure (indicating AMI) fits the data better than a simple covariance structure (indicating full invariance), without the need to specify proper priors. Several discussants highlighted merits with this approach—including its simplicity, and the limited computational resources required. The approach is being further developed to a more general class of models.

A common problem with current Bayesian approaches for measurement invariance is that they still cannot handle complex survey design (weights, stratification, clustering) easily. Complex random parameter models also have identification issues, which lead to convergence issues. When interest lies in identifying the sources of measurement non-invariance (such as the most problematic groups and items), some post-estimation diagnostic methods have been proposed, but their validity and reliability remains to be confirmed in simulation studies. On the other hand, when certain known features (such as writing system, level of development, climate zone) are expected to interfere with measurements in some predictable ways, this information can be incorporated in the priors used to estimate Bayesian random parameter models.

Discussion

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

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

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

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

In the second day of the Conference, a short session was devoted to how latent class analysis (LCA) could deal with issues of non-invariance of measurements, as they arise in large-scale international surveys. Latent class analysis refers to a class of models where both the observed responses and the unobserved trait are categorical variables (either ordinal or nominal). While this characteristic makes the models computationally demanding, treating the variable of interest as a categorical construct is often justified conceptually. In the context of large-scale international surveys, multi-group latent class analysis with ordinal classes (e.g. individuals at high, medium, or low risk) may help focus attention only on those violations of invariance that ultimately affect the classification, ignoring inconsequential violations. Similarly, addressing the question of group invariance in the context of latent class models with nominal classes, which can be often be thought of as constellations of multiple traits (e.g. personality traits, political opinions), can be less demanding than addressing this question for each of the underlying continuous traits. It is often the only alternative: in large-scale surveys, these underlying traits are often measured by short scales, if not by individual items.

The generic definition of invariance as a conditional independence property of the measurement model does also apply to latent class models; it implies that conditional on (latent) class membership, response probabilities for the observed categories do not depend on group (e.g. country) membership. In generic latent class models, where the classes are treated as nominal, the different levels of invariance (configural, metric, and scalar) do not have a clear equivalent; in contrast, different levels of invariance can be defined for latent class models in which classes are ordered (Kankaraš, Vermunt, & Moors, 2011).

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

A simple strategy to conduct LCA on international datasets is described by Eid and Diener (2001) and by Kankaraš, Moors, and Vermunt (2018). This can be described as a "bottom-up" approach: it starts by fitting country-specific latent class models in exploratory mode to find the number of classes that are supported in each country; results are then reviewed to check if all or some of the classes reflect similar patterns in responses across multiple countries.

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

A second strategy, which can be described as "top-down", is better suited for international surveys with dozens of countries (Eid, 2019). In this strategy, the exploratory step to determine the optimal number of classes is conducted directly on the pooled dataset, assuming, in a first step, that only class membership (and not country membership) determines the response patterns, while group membership only influences the size of classes. An inspection of the results can provide useful information about whether the latent classes are present in all countries. Measurement non-equivalence can manifest itself, for example, by some classes that are only present in some countries (size equal to 0). If this or other reasons (such as translation issues, different social desirability contexts) lead practitioners to suspect measurement non-equivalence, a model that allows responses to reflect not only class but also group membership would have to be specified; the more general model,

however, would have a very large number of parameters and, if group size is small, result in unstable parameter estimates. A possible solution to both issues is to specify a multi-level latent class model, where countries themselves are conceptualised as the expression of some latent set. While it may appear artificial to apply multi-level modelling to countries (which are not randomly selected groups from some overarching population, in direct violation of one of the model's assumption), treating countries as a random factor can reveal interesting sets of countries which share a common culture or institutional setting, which manifests itself in survey responses.

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

- Conducting LCA in exploratory mode on large datasets can be very time-consuming.
- Proper identification of latent classes and conditional response probabilities requires large samples (both overall, and at the group level in multi-group LCA).
- Pure statistical criteria, such as fit indices, do not always provide conclusive guidance regarding model selection, which must also be informed by priors and by qualitative judgements informed by the solution.
- Convergence issues are quite frequent with complex LCA models; default starting values may not be sufficient, and in this application, the search for the optimal number of level 2 classes (country sets) was interrupted because of convergence issues when more than 6 classes (for 38 countries) were specified for the solution.

The discussant in this session, Jeroen Vermunt, highlighted the application of LCA to surveys potentially affected by in-equivalence as an area of rapid methodological development. He also situated multi-level LCA within a more general class of multi-level mixture models, which are a way of dealing with heterogeneity by modelling the responses as reflecting different measurement models, with each model specific to a latent class of individuals or groups (e.g. countries). Among the most interesting recent contributions to the field, he singled out the development of local fit indices for multi-level LCA (Nagelkerke, Oberski, & Vermunt, 2016) and the extension of tools for quantifying the substantive impact of violations of invariance assumptions to categorical latent variables (Oberski, Vermunt, & Moors, 2015).

Improving the design of questionnaires for greater comparability of responses

The final session of the conference shifted the focus of the discussion from *statistical approaches* to *test* for invariance to *measurement approaches* that could improve the cross-country comparability of the measured constructs. When Likert scales fail to produce comparable measures of the latent construct of interest because different respondents interpret and use the response scale in a different way, a possible solution is to use different item formats and/or statistical procedures that can detect, control, and correct for differences in response styles (He et al., 2017). Indeed, no statistical procedure can remedy what a scale, by design, cannot deliver.

The discussion revolved in particular around anchoring vignettes and situational judgement tests, two innovative item formats that some large-scale surveys experimented with as a way to improve (cross-cultural) comparability. Anchoring vignettes intend to overcome the subjective nature of response scale by asking respondents to report not only a self-assessment on the scale, but also, on the same scale, how they would assess several hypothetical individuals, presented in short vignettes (King, Murray, Salomon, & Tandon, 2004). Situational judgement test items (SJT) present respondents with hypothetical situations and ask them to report "how likely" they are to act in certain ways. In this sense, they are more geared towards capturing behaviour rather than opinions. Such behavioural tendencies can be reported either on Likert-type format ("very likely", "somewhat likely", etc.) or as forced choices (e.g. by selecting the "most likely" and "least likely" options). SJTs are relatively common tools in the field of human resource management, especially in the USA. In the context of large scale international surveys, they have been recently advocated as a way to reduce social desirability bias and improve validity. They have been used in PISA 2012 (OECD, 2014) and are currently under testing for other OECD surveys.

Jonas Bertling analysed the use of anchoring vignettes and SJTs in PISA 2012 (Kyllonen & Bertling, 2013). Pauline Slot and Trude Nilsen showed how SJTs are being used in the TALIS Starting Strong Survey (TALIS-3S), aimed at pre-school educators. Jia He presented a comparative overview of different methods to improve comparability, including anchoring vignettes, direct assessments of response styles, ipsatisation (i.e. within-subject standardisation of the scores), and item parcelling (whereby individual items are combined in a single score, which is then used as indicator of a latent factor).

The three presentations highlighted the rationale for using these item types and the practical choices that need to be made when analysing the responses and reporting them on a scale.

Overall, reservations were expressed on the use of anchoring vignettes. He et al. (2017) showed that the assumption of vignette equivalence (i.e. that vignettes are understood by all respondents in the same way) is often hard to meet. In practice, the ratings observed for the hypothetical individuals often violate the expected ratings, particularly in low-ability groups (perhaps due to respondent disengagement, or to the high cognitive load that this procedure imposes to participants). The analysis of data from PISA 2012 presented by Jonas Bertling showed that accounting for vignette ratings helped aligning the between-country relationship with the within-country relationship between the latent construct and cognitive outcomes. However, some participants pointed out that these purported "improvements" in reliability and validity may be artificial, simply due to mathematical properties of the method, rather than to the substantive information gained about the response style of individuals (von Davier, Shin, et al., 2018). Furthermore, when multiple vignettes are (or could be) included in questionnaires, the choice of which vignettes to use for adjusting responses appears not to be neutral with respect to the substantive conclusions (Stankov, Lee, & von Davier, 2018).

More optimism was expressed about the potential of SJTs. The SJT items administered in the field trial of the TALIS-3S achieved scalar invariance when administered as Likert scales.

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

In his concluding comments on the studies presented in the session, Pat Kyllonen stressed the point that significant gains in comparability of survey responses across groups of respondents can also be made by following simple and universal design principles, which are not always met in practice: write items that are clearer, more concrete, behavioural, simpler, and less abstract. Overall, a strong consensus emerged around these important principles. Many participants also stressed in this occasion a point already raised repeatedly in the presentations of the first day and advised against the use of short scales (of only 3–4 items) in situations where non-equivalence at the item level is a possibility. This may well mean that rather than including many constructs, future surveys should include fewer, but better and longer scales.

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

Conclusion

The conference successfully stimulated an exchange between leading academic experts in cross-cultural measurement, industry representatives involved in the production of large-scale, cross-national survey data, and secondary users of these datasets. Opportunities to discuss these issues with a broad set of experts from different (albeit related) professions and disciplines are rare, and this exchange was highly appreciated by conference participants.

The conference was also a precious occasion for the OECD to reflect on possible improvements to the way data collected in large-scale international surveys are analysed, and how the results of the analysis are communicated to the public and to policy-makers. In this respect, two take-away messages can be drawn. First, there is a need to better communicate on issues around data comparability in OECD reports. Future technical reports should provide more extensive documentation on measurement invariance issues. In reports analysing survey results, which are targeted to a much broader audience, the challenge will be to communicate in simpler terms the complexity of the issues and the caveats that surround the analysis, at the same time extracting reasonably robust results that policy-makers could rely on. Second, a few important design principles should be better taken into account in the preparation of future survey cycles: ensure that items are crafted in a clearer, more concrete, and less ambiguous manner and prefer fewer well-crafted questions with more items over many questions with few items; produce fewer but better scales; and select the scaling method depending on the targeted concept and the reporting needs.

Additional file

Additional file 1: Annex A. Conference agenda. (DOCX 31 kb)

Abbreviations

AMI: Approximate measurement invariance; DIC: Deviance information criteria; DIF: Differential item functioning; EFA: Exploratory factor analysis; IRT: Item-response theory; LCA: Latent class analysis; MGCFA: Multi-group confirmatory factor analysis; OECD: Organisation for Economic Co-operation and Development; PIAAC: Programme for the International Assessment of

Adult Competencies; PISA: Programme for International Student Assessment; PPP: Posterior predictive *p* values; SJT: Situational judgement test; TALIS: Teaching and Learning International Survey; TALIS-3S: TALIS Starting Strong Survey

Acknowledgements
We would like to thank the many people who contributed to the success of the conference. Scientific guidance on the organisation of the conference was provided by Eldad Davidov, Michael Eid, Jean-Paul Fox, Bart Meuleman, Rens van der Schoot, and Fons van de Vijver. Miloš Kankaraš contributed to the preparatory work that led to the organisation of the conference. This meeting report benefitted from the inputs and comments of Zsuzsa Bakk, Janine Buchholz, Michael Eid, Tomoya Okubo, Artur Pokropek, and Leslie Rutkowski. The authors are solely responsible for all remaining errors and imprecisions. The opinions expressed and arguments employed in this article are those of the authors and do not necessarily represent the official view of the OECD or of its member countries.

Authors' contributions
The manuscript has been initially drafted by FA and then read, commented, and revised by MP and NLD. The authors have jointly organised the conference. The authors read and approved the final manuscript.

Funding
The authors have not received funding.

Availability of data and materials
Not applicable. No data are used in the paper.

Competing interests
The authors declare that they have no competing interests.

Received: 19 February 2019 Accepted: 12 July 2019
Published online: 06 August 2019

References

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210.

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49*(5), 713–734. https://doi.org/10.1177/0022022117749042.

Buchholz, J., & Hartig, J. (2017). Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Applied Psychological Measurement*. https://doi.org/10.1177/0146621617748323.

Byrne, B., Shavelson, R., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456.

Davidov, E., & Meuleman, B. (2019). Measurement invariance analysis using multiple group confirmatory factor analysis and alignment optimisation. In F. J. van de Vijver (Ed.), *Invariance analyses in large-scale studies*. Paris: OECD Publishing. https://doi.org/10.1787/254738dd-en.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*(1), 55–75. https://doi.org/10.1146/annurev-soc-071913-043137.

De Jong, M., Steenkamp, J.-B., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research, 34*(2), 260–278. https://doi.org/10.1086/518532.

Eid, M. (2019). Multigroup and multilevel latent class analysis. In F. J. van de Vijver (Ed.), *Invariance analyses in large-scale studies*. Paris: OECD Publishing. https://doi.org/10.1787/254738dd-en.

Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology, 81*(5), 869–885. https://doi.org/10.1037/0022-3514.81.5.869.

Fox, J.-P. (2019). Cross-cultural comparability in questionnaire scales: Bayesian marginal measurement invariance testing. In F. J. van de Vijver (Ed.), *Invariance analyses in large-scale studies*. Paris: OECD Publishing. https://doi.org/10.1787/254738dd-en.

Fox, J.-P., Mulder, J., & Sinharay, S. (2017). Bayes factor covariance testing in item response models. *Psychometrika, 82*(4), 979–1006. https://doi.org/10.1007/s11336-017-9577-6.

Haberman, S. (2009, *2009*). Linking parameter estimates derived from an item response model through separate calibrations. *ETS Research Report Series*, (2), i-9. https://doi.org/10.1002/j.2333-8504.2009.tb02197.x.

He, J., Van de Vijver, F., Fetvadjiev, V., de Carmen Dominguez Espinosa, A., Adams, B., Alonso-Arbiol, I., et al. (2017). On enhancing the cross-cultural comparability of Likert-scale personality and value measures: A comparison of common procedures. *European Journal of Personality, 31*(6), 642–657. https://doi.org/10.1002/per.2132.

Horn, J., & Mcardle, J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117–144. https://doi.org/10.1080/03610739208253916.

Kankaraš, M., Moors, G., & Vermunt, J. (2018). Testing for measurement invariance with latent class analysis. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-cultural analysis : Methods and applications*. Routledge. https://doi.org/10.4324/9781315537078.

Kankaraš, M., Vermunt, J., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research, 40*(2), 279–310. https://doi.org/10.1177/0049124111405301.

King, G., Murray, C. J., Salomon, J. A., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review, 98*(01), 191–207. https://doi.org/10.1017/S000305540400108X.

Kyllonen, P., & Bertling, J. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment : Background, technical issues, and methods of data analysis*. Boca Raton: Chapman and Hall/CRC.

Lek, K., & van de Schoot, R. (2019). Bayesian approximate measurement invariance. In F. R. van de Vijver (Ed.), *Invariance analyses in large-scale studies*. Paris: OECD Publishing. https://doi.org/10.1787/254738dd-en.

Lommen, M., van de Schoot, R., & Engelhard, I. (2014). The experience of traumatic events disrupts the measurement invariance of a posttraumatic stress scale. *Frontiers in Psychology, 5*, 1304. https://doi.org/10.3389/fpsyg.2014.01304.

Mellenbergh, G. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825.

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: The alignment method. *Frontiers in Psychology, 5*, 978. https://doi.org/10.3389/fpsyg.2014.00978.

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups. *Sociological Methods & Research, 47*(4), 637–664. https://doi.org/10.1177/0049124117701488.

Muthén, L., & Muthén, B. (1998-2017). *Mplus User's Guide* (Eighth Edition ed.). Los Angeles: Muthén & Muthén.

Nagelkerke, E., Oberski, D., & Vermunt, J. (2016). Goodness-of-fit of multilevel latent class models for categorical data. *Sociological Methodology, 46*(1), 252–282. https://doi.org/10.1177/0081175015581379.

Oberski, D., Vermunt, J., & Moors, G. (2015). Evaluating measurement invariance in categorical data latent variable models with the EPC-interest. *Political Analysis, 23*(04), 550–563. https://doi.org/10.1093/pan/mpv020.

OECD. (2014). *PISA 2012 Technical Report*. Paris: OECD Publishing Retrieved 06 27, 2019, from http://www.oecd.org/pisa/data/pisa2012technicalreport.htm.

OECD. (2017). *PISA 2015 Technical Report*. OECD publishing Retrieved 11 27, 2017, from http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf.

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–21. https://doi.org/10.1080/10705511.2018.1561293.

Robitzsch, A., & Lüdtke, O. (2018). Linking errors in international large-scale assessments: Calculation of standard errors for trend estimation. *Assessment in Education: Principles, Policy & Practice*, 1–22. https://doi.org/10.1080/0969594X.2018.1433633.

Rutkowski, L., & Svetina, D. (2013). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and*

Psychological Measurement, 74(1), 31–57. https://doi.org/10.1177/0013164413498257.

Rutkowski, L., & Svetina, D. (2016). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education, 30*(1), 39–51. https://doi.org/10.1080/08957347.2016.1243540.

Simmons, J., Nelson, L., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632.

Stankov, L., Lee, J., & von Davier, M. (2018). A note on construct validity of the anchoring method in PISA 2012. *Journal of Psychoeducational Assessment, 36*(7), 709–724. https://doi.org/10.1177/0734282917702270.

Steenkamp, J.-B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78–107. https://doi.org/10.1086/209528.

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 770. https://doi.org/10.3389/fpsyg.2013.00770.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740.

von Davier, M., Shin, H.-J., Khorramdel, L., & Stankov, L. (2018). The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement, 42*(4), 291–306. https://doi.org/10.1177/0146621617730389.

von Davier, M., Yamamoto, K., Shin, H.-J., Chen, H., Khorramdel, L., Weeks, J., et al. (2018). Evaluating item response theory linking and model fit for data from PISA 2000–2012. In *Assessment in Education: Principles, Policy & Practice* Special Issue.

Publisher's Note