A Cross-Cultural Basis for Public Service? Public Service Motivation Measurement Invariance in an Original Survey of 23,000 Public Servants in Ten Countries and Four World Regions

Kim Sass Mikkelsen, Roskilde University

Christian Schuster, University College London

Jan-Hinrik Meyer-Sahling, University of Nottingham

This is the peer reviewed version of the article that has been published in final form by the 'International Public Management Journal'. Tables, figures and appendices are available upon request. This paper received support from the Global Integrity Anti-Corruption Evidence (GI-ACE) programme, funded with UK aid from the UK government.

Abstract

Public service motivation (PSM) is a core concept in public administration, studied in surveys across numerous countries. Whether these studies accumulate comparable knowledge about PSM crucially depends on PSM measurement invariance: that PSM has a similar measurement structure in different national contexts. Yet, large-scale cross-country research to address this conundrum remains scant. Drawing on an original survey of 23,000 public servants in ten countries in Eastern Europe, Asia, Latin America, and Africa, our paper addresses this gap. Replicating Kim et al.'s 16-item scale, we find partial metric invariance for the four PSM dimensions in eight countries, but scalar non-invariance. This suggests that results from structural equations about the causes and consequences of PSM may be compared across most countries, yet means of PSM and its dimensions are not generally comparable. PSM research thus cannot adjudicate in which countries public service motivation is higher or lower on average but can compare relationships between PSM and individual characteristics or management practices between countries. Our findings underscore the cross-cultural basis of public service motivation and its limits.

1 Introduction

Among topics in public administration, public service motivation (PSM) research "stands out by [its] sheer numbers", with more than fifty studies published annually in the last years (Ritz, Brewer, and Neumann 2016; Rainey and Steinbauer 1999, 20). PSM is typically understood as a "particular form of

altruism or prosocial motivation that is animated by specific dispositions and values arising from public institutions and missions" (Perry and Hondeghem 2008, 3). PSM research has been instrumental in advancing our understanding of how to motivate public employees – one of the 'big questions' in public management (Behn 1995, 313). Public managers often have less leverage over other motivators – such as performance incentives – putting a premium on leveraging PSM as an alternative source of work motivation in the public sector (Esteve and Schuster 2019). PSM research offers a range of practical insights to this ends (Christensen, Paarlberg, and Perry 2017). The globalization of PSM research – with PSM studies increasingly conducted across world regions – implicates that these insights are usefully drawn from an increasingly diverse set of contexts (Ritz, Brewer, and Neumann 2016).

These inferences of PSM research are overwhelmingly based on individual-level survey measures of PSM (Ritz, Brewer, and Neumann 2016). Respondents are asked to indicate the extent of their agreement with measures such as "I am prepared to make sacrifices for the good of society" (Kim et al. 2013). Measurement scales often cover several – and typically four – PSM dimensions such as self-sacrifice and compassion.

In light of 1) the centrality of PSM research for the scholarly and practitioner understanding of the nature of bureaucracy and public service around the world and 2) the hundreds of PSM studies across the globe relying on PSM survey scales, one would expect a large industry of scholarship that rigorously assesses whether PSM measures are comparable across different contexts and countries. Cross-country comparability of survey measures is anything but a foregone conclusion. Comparability is likely not helped by the potentially culturally loaded content of many PSM survey items. To cite just two illustrative items from PSM measurement scales: what meaningful public service ("Meaningful public service is very important to me") or civic duty ("I believe in putting civic duty before self") means to respondents may well vary across cultural contexts and threaten comparability of measures and conclusions.

Without evidence on cross-cultural measurement invariance - comparability of latent measurement scales across cultures - knowledge accumulation in PSM research is heavily impaired. If two PSM studies in two different countries found diverging effects of PSM, for instance – as systematic literature reviews frequently suggest (Ritz, Brewer, and Neumann 2016) – it would remain altogether unclear whether that would implicate that PSM has different substantive effects in the two countries – or if public service motivation

indicators simply measure their latent scales differently in one country than in another. If PSM measurement differs in different cultures and languages in turn, generalizations about PSM would scarcely be possible.

This would also implicate that highly-cited systematic literature reviews of PSM – which sum up studies finding positive or negative effects of PSM across countries (e.g. Ritz, Brewer, and Neumann 2016) – may provide invalid insights, as might meta-analyses of the causes and consequences of PSM across studies and countries (e.g. Harari et al. 2016; Awan, Bel, and Esteve 2018). Similarly, the validity of inferences from PSM studies focused on comparing PSM levels across countries may be in doubt (e.g. Vandenabeele and Van de Walle 2008).

In other words, systematic cross-cultural and cross-national measurement invariance analyses are central to gauge the comparability and generalizability of the large body of substantive PSM findings, and to enable meaningful knowledge accumulation in PSM research. Despite that, quantitative PSM research has been largely mute about them. The only significant exception is (Kim et al. 2013). Kim et al. (2013) drew their inferences from a sample of a total of 2,868 local government employees in 12, mostly Western European, countries. While Kim et al.'s (2013) study crucially expands our understanding of cross-national measurement equivalence of PSM, it falls short of providing a conclusive answer. Two shortcomings – both of which our paper addresses – stand out.

The first is methodological. Kim et al. (2013) tested only full metric invariance of their 16-item battery, constraining all factor loadings to be equal across all countries. Based on this, Kim et al. (2013) find violations of metric invariance. Yet, this benchmark is rarely met in international survey research on any topic (see e.g. Davidov, Schmidt, and Billiet 2018) and often need not be met for acceptable comparisons of estimates across countries. The literature on measurement invariance in cross-cultural research, instead, recommends a different standard: partial metric equivalence. A typical recommendation is that at least two item loadings must be equal for a latent variable to display metric equivalence. Our paper follows this second, more widely-accepted approach in international survey research.

The second limitation is empirical. While Kim et al.'s (2013) sample is impressive, it is heavily tilted towards Western Europe and countries of the Global North. Entire world regions – such as Africa and Latin America – are missing from the sample. Whether PSM has a similar measurement scale across these regions thus remains unclear. Moreover, Kim et al.'s (2013) sample size in each country is relatively small, with an

average of 239 respondents per country; and the sample is drawn from a convenience sample of local governments. The small, unrepresentative samples risk type II errors about measurement invariance.

In defense of Kim et al. (2013), collecting a larger and more representative cross-country survey sample of public servants is time- and funding-intensive, and a serious logistical challenge. It requires original survey administration across multiple languages and countries, with access to a larger number of government employees and organizations. These barriers may well explain why PSM researchers have – notwithstanding limited evidence on their cross-country measurement invariance – not prioritized undertaking a large-scale cross-country PSM survey to understand whether findings across countries may be compared in the first place. This reflects a more general dearth of cross-national equivalence analyses of measurement scales in public administration research – despite recent calls to strengthen comparative public administration (Fitzpatrick et al. 2011).

Drawing on a large-scale original survey data collection effort with 23,000 central government employees in ten countries and several hundred government institutions in Eastern Europe, Asia, Latin America and Africa – the largest full-scale PSM sample in the literature to-date – our paper addresses these gaps (included countries are Albania, Estonia, Kosovo, Bangladesh, Nepal, Brazil, Chile, Ghana, Malawi, and Uganda). It provides an empirical basis for understanding cross-country and cross-cultural PSM measurement invariance on a much larger scale than the only previous effort (Kim et al. 2013). It comes with a survey sample that is eight times as large and spanning, in each country, across a broader range of government institutions. As such, it provides an empirical foundation for claims to comparability and knowledge accumulation across countries in PSM research.

Our results are, overall, good news for PSM research in public administration. Replicating Kim et al.'s (2013) 16-item PSM scale, we are able to show partial metric invariance for the four PSM dimensions across eight of our countries and three regions (Eastern Europe, Latin America, Africa). Our two (South) Asian cases, which showed worse model fit, were the sole exception. This should give cause for comfort in the PSM research community as it implies that results from structural equations about the causes and consequences of PSM may, in fact, be reasonably compared across most cultural settings. Balkanization of knowledge can be avoided, even as PSM research goes global and enters the developing world. This also implies that the findings of systematic literature reviews and meta analyses of PSM are meaningful (Harari et

al. 2016; Awan, Bel, and Esteve 2018; Ritz, Brewer, and Neumann 2016): with partial metric invariance, the signs and size of coefficients of the causes and consequences of PSM can be compared across (most) countries.

Our findings, however, also underscore limits to the cross-cultural basis of PSM. First, we fail to uncover scalar invariance in our sample. Thus, means of PSM and its dimensions are not comparable across countries. As comparisons of means of PSM and PSM dimensions are not meaningful, cross-country PSM surveys cannot provide insights into which countries' public officials are more or less motivated to serve the public. This, unfortunately, sheds significant doubt on the validity of PSM studies which derive their inferences from comparing PSM levels across countries (e.g. Vandenabeele and Van de Walle 2008).

Second, our results suggest that the extent of measurement invariance differs across PSM dimensions. Self-sacrifice and, to a lesser extent, compassion are relatively invariant. By contrast, commitment to public values and attraction to public service are more non-invariant. The two most "public" dimensions of PSM are thus most affected by measurement non-invariance. While troubling, this is intuitively plausible: public values, for instance, may differ across different national settings – and so does the meaning of commitment to public values.

In sum, our findings suggest that (1) the PSM measurement battery developed by Kim et al. (2013) – which our paper validates with large samples in eight of ten surveyed countries – is a solid measurement tool for future PSM research in most, but not all countries. (2) That, contrary to the conclusion in Kim et al. (2013), PSM structural regression estimates are comparable across even very different countries; knowledge accumulation in PSM research across national settings, including through meta analyses, is thus feasible. (3) However, country-level comparisons of PSM levels are not valid due to scalar non-invariance. (4) Finally, our findings raise important questions for future measurement research about non-invariance of the more "public" dimensions of PSM: commitment to public values and attraction to public service.

The paper proceeds as follows. We, first, outline the theory behind measurement invariance testing. In particular, we discuss the standard multi-group confirmatory factor analysis (MGCFA) framework and briefly debate some alternatives. Thereafter, we explain our model building, estimation, and validation strategy. Subsequently, we discuss our PSM measurement and our survey sample. This is followed by a

discussion of our results, proceeding from configurational, through first-order and second-order metric, to scalar invariance tests. Finally, we discuss the consequences of our results for PSM research and conclude.

2 Measurement Invariance

Since the late 1990s, a standard approach to measurement invariance in a MGCFA framework has developed encompassing configurational invariance, metric invariance, scalar invariance, and strict invariance (see, among many Davidov, Schmidt, and Billiet 2018; Putnick and Bornstein 2016). These types of invariance are viewed as hierarchically organized, with each higher order of invariance assuming all lower orders.

Configurational invariance denotes equivalence of model form, requiring simply that the same factor structure be modelled across groups. Without configurational invariance, no meaningful comparisons across groups are possible. Failing model fit in some groups, configurational invariance could be compromised if different models were to be estimated in different groups. Alternatively, a model search could begin to find a model that fits all groups. For our purposes, this is not feasible as we strive to test an established four-factor structure rather than questioning it.[1]

Metric invariance requires, in addition to configurational invariance, that an equality constraint be imposed on factor loadings across groups. This ensures that structural regression estimates are comparable across groups. Without metric invariance, the sign of these estimates are comparable across groups but effect sizes are not.

Scalar invariance requires, in addition to metric invariance, that an equality constraint be imposed on item intercepts across groups. This ensures comparisons of latent means are comparable across groups. Without it, group specific answers to items prevents meaningful comparisons of means.

Finally, *strict invariance* requires, in addition to scalar invariance, that variances are equal across groups. This is useful chiefly if variances are of substantive interest.

Since PSM is frequently considered a second-order latent construct, wherein survey items relate to the four dimensions which in turn relate to PSM, it is necessary to consider first and second-order invariance. The two-level structure creates an additional complication for measurement invariance as the invariance of second-order factor loadings and intercepts depends on first-order invariance. We follow the

recommendation by Chen, Sousa, and West (2005) that first-order invariance be established before second-order invariance is tested. Like types of invariance, invariance of orders are hierarchically organized. Hence, after establishing configurational invariance, we test first-order metric invariance, followed by second-order metric invariance, followed by first-order scalar invariance, followed by second-order scalar invariance.

Between both types of invariance and orders of constructs, invariance is tested in the MGCFA framework using model comparisons. Metric invariance is tested through a comparison of the fit for a model including equality constraints on factor loadings across groups with a model imposing no such constraints. The fit of the former model will be worse than the less constrained latter model. The question answered in measurement invariance testing is whether this fit deterioration is sufficiently small to be ignorable. Scalar invariance is similarly tested through a comparison of a model constraining both factor loadings and item intercepts across groups with a model constraining only factor loadings.

As noted in the introduction, full metric and scalar invariance is rare in cross-cultural research. Consequently, researchers frequently apply partial invariance procedures to test their constructs (Davidov, Schmidt, and Billiet 2018). Partial invariance approaches constrain some but not all items when testing whether constructs are invariant. If a sufficient number of item loadings or intercepts – typically a majority or two per construct – can be constrained without a substantial deterioration of model fit, the model is considered as featuring partial metric or scalar invariance respectively. This is the approach we take to our data.[2]

3 Model building, testing, and identification approach

For our estimates, we rely on the cfa function from the lavaan package for R (Rosseel 2012). Since our observed variables will be ordered categorical answers to survey items - and since some variables show signs of skew - we use a robust version of diagonally weighed least squares (DWLS) as our estimator, and robust fit measures. In the remainder of this section, we discuss the choice of fit measures for model comparisons, our strategy for identifying latent variables, and our approach to testing partial measurement invariance while avoiding sample specific model building.

3.1 Fit measures and benchmarks

The most common benchmark for testing measurement invariance in the literature is likely the ΔCFI. Cheung and Rensvold (2002) proposed to reject measurement invariance if ΔCFI < -0.010.[3] A similar 0.010 benchmark for ΔRMSEA has been suggested in the literature but is less established (see, e.g. Rutkowski and Svetina 2014; Davidov, Schmidt, and Billiet 2018). Finally, a significance benchmark exists for $\Delta\chi^2$, as differences in this fit index can be statistically tested. As is common in the literature, we do not rely on this measure. There are three reasons for this choice. First, $\Delta\chi^2$ does not follow a χ^2 distribution when robust versions of the fit index are used. Second, with large datasets such as ours, significance testing will tend to over-reject invariance as differences may be statistically significant but substantially irrelevant. Third, $\Delta\chi^2$ depends on the fit of the unrestricted model in ways that ΔCFI and ΔRMSEA do not (Yuan and Bentler 2004).

Consequently, we rely primarily on ΔCFI as our primary benchmark while we report ΔRMSEA and $\Delta\chi^2$ for reference and without reporting a significance test of the latter. With respect to benchmarks, we rely primarily on the -0.010 benchmark for ΔCFI, supported by the 0.010 threshold for ΔRMSEA.

It is worth noting that the application of those standard benchmarks to invariance testing across many groups has seen significant discussion in the literature. Rutkowski and Svetina (2014) recommend based on a simulation that more liberal thresholds for metric invariance testing (-0.020 for ΔCFI and 0.030 for ΔRMSEA) be used for large numbers of groups (20 in their simulation). For scalar invariance testing they recommend the standard thresholds. However, in a simulation with 10 groups, as in our setting with 10 countries, they find standard benchmarks to be able to discriminate satisfactorily between metric invariance and non-invariance. Hence, while measurement invariance assessments with more groups than our ten countries may utilize more lenient benchmarks for metric invariance testing, we opt for the standard benchmarks rather than risk inferences based on a benchmark that may be too lenient.[4]

3.2 Identification

The most common approach to giving scale to dimensions of PSM is using a marker variable strategy, fixing the factor loading of one item per dimension to unity to give scale to the respective latent variable. This is sensible in general, but is not ideal for testing measurement invariance of multidimensional constructs. Configurational invariance requires that the same estimation strategy be used across groups. Consequently, the loading of one item for each dimension – the loading of the marker variable – features metric invariance by design. A similar point applies to scalar invariance as the identification of latent means in the marker variable strategy requires fixing the mean of the marker variable to zero, generating scalar invariance for that variable by design. In the literature, as a result, the choice of marker variables is a focal point, since the use of a marker variable that is not invariant will tend to reject invariance in instances where it does in fact hold (Davidov, Schmidt, and Billiet 2018). We could have taken a theoretical approach to this problem, or a data driven one and probed which item from each dimension provides the best result. However, evaluation of invariance might still be influenced by the choice of marker variables.

To avoid this issue altogether, we instead opted to give scale to our latent variables using Little et al.'s (2006) effects coding strategy. In this framework, latent variables are given scale by constraining the average of their item loadings to unity and the sum of their means to zero. The result, for our purpose, is twofold. First, latent variables retain the scaling of their indicators. Second, as no marker variable is used we are not constraining any loadings or intercepts to be equal across groups by design.

3.3 Approach to partial invariance

One obvious problem with partial invariance models is which loadings or intercepts should be constrained to be equal across groups. Our solution is to use a split-sample validation strategy for model building and testing. Within each country, we randomly divide respondents into a training dataset and a validation dataset. Subsequently, we identify the best fitting partial invariance model in the training data and subsequently implement it on the validation data. In this way, we are able to demonstrate that our conclusions are not sample specific through validation.[5]

How, then, do we determine which constraints should be loosened? We follow Lee et al.'s (2018) approach and consider differential item functioning (DIF) across our countries.[6] In particular, upon rejecting full

metric or scalar invariance, we utilize a free-baseline strategy: (1) loosening all relevant equality constraints in the model (e.g. all factor loadings), (2) reiteratively imposing equality constraints one item at a time, and (3) evaluating deterioration in model fit for each constrained item. Items that result in deteriorated fit are determined to have DIF and should not be constrained in the partial equivalence model.[7] We exclude restrictions on the items that have the largest deterioration. Partial invariance obtains so long as each dimension of the PSM construct - and the PSM construct itself at the second-order level - causes at least two variables (or dimensions), which do not display DIF.

3.4 Overview of analyses

Our strategy, in sum, follows several steps in sequence (figure 1). Starting with the training data, we fit the same model to all countries to ensure configurational invariance and to test model fit within each country. Subsequently, we fit metric invariance restrictions at, initially, the first and then at the second-order level of the PSM construct. After that, we fit scalar invariance restrictions at, initially, the first and then at the second-order level of the construct. Finally, we assess whether the model we have built shows invariance in the validation data.

If the model does not show configurational invariance, the analysis ends there. If configurational invariance obtains, we, first, test first full metric invariance and, failing that, partial metric invariance. If neither type of metric invariance obtains, we simply test whether the model also fits in the validation data and end the analysis. If either full or partial metric invariance holds, we test for full and, failing that, partial second-order metric invariance. If the data supports neither full nor partial second-order metric invariance, we test our model for full or partial first-order metric invariance in the validation data and end the analysis. If the data supports either full or partial second-order metric invariance, we repeat the process for first- and second-order scalar invariance. If neither is supported in the training data, we test our model for first and second order metric invariance with the validation data. Our evaluation of first- and second order scalar invariance follows a similar logic as shown in the figure.

[Figure 1 around here]

4 The PSM Construct and Measurement

Perry (1996) originally built a PSM construct consisting of four dimensions: commitment to the public interest, compassion, self-sacrifice, and attraction to policy making. Subsequent multidimensional research has attempted, with some exceptions, to retain a four-factor structure (Ritz, Brewer, and Neumann 2016). Quite a few applications now replace attraction to policy making with attraction to public service and commitment to the public interest with commitment to public values (e.g. Kim et al. 2013; Meyer-Sahling, Mikkelsen, and Schuster 2017). While there is still some debate concerning the right factor structure and the discriminant validity of factors (see e.g. Kim et al. (2013)), the majority of studies reviewed recently by Ritz et al. (2016) followed one of the two four-factor models. Hence, compassion, self-sacrifice, commitment to public values or interests, and attraction to public service or policy now are at the heart of multidimensional PSM constructs.

In our analysis, we aim to support this practice by evaluating measurement invariance for Kim et al.'s (2013) four-factor model. We chose to rely on Kim et al.'s (2013) scale, both as Kim et al.'s (2013) dimensions are considered as the "current authority" in at least some recent works (Prebble 2016, 268), and as, to our knowledge, Kim et al.'s (2013) scale is the only one which has undergone a prior cross-country measurement invariance exercise.

Table 1 lists Kim et al.'s (2013) 16 items and 4 dimensions: attraction to public service (APS), commitment to public values (CPV), compassion (COM), and self-sacrifice (SES).

[Table 1 around here]

Deciding on the number and content of dimensions, of course, does not in itself answer the question how these dimensions relate to the overarching PSM construct. In the CFA and SEM frameworks that applied PSM research frequently applies, it seems natural to model PSM as a reflective second-order latent construct, in which PSM causes its dimensions, which in turn cause their indicators. Researchers have made the argument that this is the correct way of specifying the construct and some applications do model PSM using this approach (e.g. Clerkin and Coggburn 2012; Meyer-Sahling, Mikkelsen, and Schuster 2017). We follow this approach in our empirical analysis.

This is, of course, not the only modelling strategy. To get an overview of strategies and make our analysis as consistent with the literature as possible, we conducted a review of modelling choices in 97 published PSM studies (see Appendices G and H). In this review, most PSM studies either do not consider a second-order construct at all or construct PSM as a composite directly from dimensions – for instance by summing or averaging factor scores. For these studies, measurement invariance at the first order would suffice. Some studies model, as we do, PSM as a reflective second-order latent construct – and thus include a testable second-order latent construct. Barely any study we reviewed relies on the first-order reflective, second-order formative model proposed by Kim (2011). Given this lack of application in PSM research, we do not conduct separate tests for measurement invariance of PSM as a formative latent construct.

5 Survey Sample

To conduct our measurement invariance analysis, we surveyed 23,000 public servants in ten governments – to our knowledge, the largest full-scale PSM survey in the literature to-date. To ensure a diverse population of public servants to assess measurement invariance and, concomitantly, the cross-cultural basis of public service motivation, our survey sample comprises public servants across ten countries in four developing regions: Latin America (Brazil and Chile), Eastern Europe (Estonia, Kosovo and Albania), Africa (Ghana, Malawi and Uganda) and Asia (Nepal and Bangladesh). Our case selection ensures a heterogeneity of contexts, in terms of not only different regional and thus cultural contexts, but also low and high income, democratic and (partially) autocratic, and low and high corruption perception (see Appendix A).

In each country, we surveyed a comparable set of respondents: public servants in central governments across ranks (from administrative assistance to management); working in central government institutions (that is ministries and agencies, rather than municipal or state governments); and undertaking administrative functions in the broadest sense (excluding, e.g., policemen, military, teachers or doctors).[8]

While we surveyed comparable populations of public servants across countries, local contexts obliged us to rely on two distinct survey modes across countries. In our Eastern European and Latin American cases, governments counted on records of email addresses of public servants. We were thus able to conduct surveys online. In Estonia, Kosovo, and Albania, all civil servants were invited via email to respond to the survey,

except officials employed in defence ministries and their subordinated organizations. In Brazil and Chile, all civil servants in eleven central government institutions (Chile) and fourteen federal government institutions based in Brasilia (Brazil) were invited to participate in the survey. The online surveys were conducted between November 2016 and December 2017. Response rates ranged from 11% to 47% and, in total, between 2,431 and 5,742 responses were collected in each country (see Table 2).

Limitations in email records and computer access of public servants precluded similar online survey sampling in our African and Asian cases. Moreover, weak personnel records – governments do not have, or were not willing to disclose, complete lists of public employees in central government institutions – precluded strictly representative samples. As a result – and similar to a range of prior studies surveying bureaucrats in developing countries (see, e.g. Meyer-Sahling and Mikkelsen 2016; Oliveros and Schuster 2018) – we lacked the requisite survey frames for representative surveys of public servants. Instead then, we had to rely on informal quota sampling and in-person surveys.

This informal quota sampling aimed to ensure that public servants across a range of central government organizations, hierarchical levels, job functions, contract types, ages and education levels were sampled. Sampling was based primarily on contacting government organizations one-by-one and asking for access, with an effort to stratify the sample in a general sense across central government. Subsequently, local enumerators conducted in-person interviews with public servants. Between February and December 2017, our enumerators interviewed between 1,077 and 1,645 public servants per country.

In total, the survey sample included 48 (Ghana), 31 (Uganda), 62 (Malawi), 31 (Nepal) and 38 (Bangladesh) government institutions. Similarly, our online surveys included responses from 11 (Chile), 18 (Albania), 26 (Brazil), 53 (Estonia), and 83 (Kosovo) government institutions. No institution takes up more than 26.6% of a country's responses (which the Ministry of Finance and its subordinated agencies does in the Brazil sample). Table 2 provides an overview of our survey samples.

[Table 2 around here]

Our sampling strategy yielded a diverse set of public servants in each surveyed country. Respondents are roughly split on gender. They are mostly (60%) public servants working in professional ranks, though with important shares in administrative support (23%) and managerial (17%) ranks. A large majority (77%) – though far from all – are employed on permanent contracts. On average, our respondents are 43 years old, and have worked for over 13 years in the public sector.

Where we can assess representativeness thanks to data availability - Bangladesh, Brazil, Chile, Estonia, Ghana, and Uganda - we find that our samples roughly approximate our survey populations in gender (Brazil, Bangladesh, Ghana, and Chile) and age (Estonia, Chile, Brazil, and Uganda) in most countries with those demographics available. Our respondents tend to be, with the exception of Chile, somewhat more educated than average central government employees (though this stems in part from our survey samples excluding groups such as armed forces, while available government survey population data does not always do so). In four countries, government collaborators either did not have or did not share aggregate staff data or survey population data. At least based on available demographics, our survey samples in both in-person and online surveys appear to meaningfully reflect local survey populations on at least some demographics (see Appendix B), but, as noted, fall short of allowing us to make strong representativeness claims.

In each country with local languages, our PSM measures were translated from English into the local language(s). To safeguard a comparable understanding of the wording of our questions across our diverse range of countries and languages, we pre-tested our survey in each country through a series of cognitive interviews with public servants. In each country, measures were iteratively revised in local languages until cognitive interviews suggested measures were understood as intended.

Table 3 shows the descriptive statistics for the sixteen item battery in the resulting sample across all ten countries (for descriptive statistics by country see tables B4-B8 in the Appendix).

[Table 3 around here]

6 Results

In line with our methodological approach to assessing measurement invariance, we conduct increasingly demanding invariance tests: first, configurational invariance; then, first and second-order metric invariance; and, lastly, scalar invariance.

6.1 Configurational invariance

Kim et al. (2013) test configurational invariance by testing if models other than their preferred four-factor model fit the data better in their 12 countries, finding support for this in eight. This is not, in fact, required for configurational invariance to hold. Configurational invariance only requires the same model to be estimated and fit in all groups – not that this is the best performing model in all groups.[9]. We thus simply estimate the fit of the four-factor model in each country to assess configurational invariance.

Figure 2 shows the result of this analysis, giving the χ^2 contribution per respondent, as well as the CFI, and the RMSEA for each country (see Appendix C for further details).[10] We show conventional benchmarks for good and acceptable fit on the two latter indices in the figure as dotted lines (e.g. Hu and Bentler 1999; Byrne 2008). As the analysis shows, the PSM dimensions fit the data well in most countries. The only exceptions are acceptable but not good fits in our two Asian countries on the CFI and a marginally less than good fit in Estonia on the RMSEA.

[Figure 2 around here]

As we will use the CFI as the main criterion for our DIF and measurement invariance, this raises some concerns about the Asian cases. As fit deterioration will occur for every set of constraints we introduce in measurement invariance testing, less than good fits can be expected to create problems. In fact, estimating models on all ten cases does not support and validate partial metric invariance using standard benchmarks. Since the purpose of this paper is to examine the boundaries of the comparability of the PSM construct, we demonstrate below partial metric invariance in a subset of eight countries (rather than, as we find, the lack of the same in ten). We return to our Asian cases in the discussion section.

Applying the four-factor model on the remaining eight countries, we arrive at the following conclusions: The model permitting factor loadings and intercepts to vary across countries at all levels gives a good fit (χ^2 = 3740.76, df = 980, p-value < 0.001; RMSEA = 0.026; CFI = 0.983). A model with a reflective second-order

construct gives a similarly good fit (χ^2 = 4051.69, df = 1.000, p-value < 0.001; RMSEA = 0.029; CFI = 0.979).

6.2 First-order metric invariance

The test for full metric invariance returns a good fit for the fixed-loadings model (χ^2 = 3831.52 df = 868, p-value < 0.001; RMSEA = 0.032; CFI = 0.972). However, the fit deteriorates compared to the configurational model beyond Cheung and Rensvold's (2002) ΔCFI benchmark ($\Delta\chi^2$ = 740.52; ΔRMSEA = 0.008, ΔCFI = 0.013). For this reason, we proceed to examine partial metric invariance. It is worth noting that more lenient thresholds in the literature for analyses with a large number of groups would imply support for full first-order metric invariance from this analysis (Rutkowski and Svetina 2014). However, as discussed previously, our sample does not have enough groups, in our view, for these benchmarks to apply.

The first step of our model building for partial metric invariance is determining DIF for each item in our model. Following our free-baseline strategy, we restrict the factor loading one item at a time and estimate the deterioration of fit. Figure 4 shows the resulting absolute change in CFI and RMSEA for each constrained item (see Appendix D for details). Larger values implies a higher degree of DIF. This means that, if we were to obtain partial metric invariance by loosening constraints on only one item, we should choose COM3.

[Figure 3 around here]

The cost of our identification strategy becomes visible here. While we are able to avoid arbitrarily constraining four factor loadings to equality (and unity) across countries, we cannot loosen one item only. Effects coding identifies the factor by setting the average loading to unity for each factor, which means loosening only one loading will result in an equal estimate across countries due to the identification constraint in spite of being free across countries. Hence, to let COM3 be estimated freely across countries, we need to let another item reflecting COM also be freely estimated. Inspection of figure 3 will show that COM2 is the best candidate for a pair, since it is the measure of COM that results in the second-largest fit deterioration when constrained.

To test whether releasing constraints on COM2 and COM3 is sufficient to obtain partial metric equivalence, we fit a model constraining all factor loadings except COM2 and COM3 to be equal across countries. This model fits the data well (χ2 = 3637.22, df = 861, p-value < 0.001; CFI = 0.975; RMSEA = 0.031) but still falls just short of the benchmark for invariance ($\Delta\chi^2$ = 546.22; ΔCFI = 0.011; ΔRMSEA = 0.006).

While we could accept this deterioration in global fit measures as acceptable, acknowledging that the 0.01 benchmark is not a hard distinction between acceptable and non-acceptable, we proceed to a second round of DIF testing. We constrain COM1 and COM4 to be equal across countries as COM2 and COM3 are freely estimated and two items are required per dimension for partial metric equivalence. Subsequently, we estimate a model constraining each item in APS, CPV, and SES reiteratively and see which constraint deteriorates fit the most relative to the COM-constrained model.

[Figure 4 around here]

This analysis, illustrated in figure 4, singles out CPV1 and CPV2 as the best candidates for DIF. Consequently, the next step is loosening factor loadings for these items, along with COM2 and COM3, while fixing CPV3 and CPV4, along with COM1 and COM4. The resulting model fits the data well (χ^2 =3431.06, df = 854, p-value < 0.001; CFI = 0.977; RMSEA = 0.030). Moreover, it does not permit rejection of partial metric measurement invariance along conventional benchmarks ($\Delta\chi^2$ = 340.06; ΔCFI = 0.008; ΔRMSEA = 0.005).

While we could end our DIF analysis here based on global fit measures, we proceeded to perform a third round of DIF testing to examine if any of the remaining dimensions, APS and SES, show signs of DIF comparable to what our analysis revealed for COM and CPV. In particular, from figures 3 and 4, it appears that items APS1 and APS4 contribute about as much to fit deterioration as COM2, which we do not constrain as a consequence of our previous analyses. The assumption in partial measurement invariance testing is that any constrained loading has ignorable DIF. From this perspective, small deterioration in global fit indices may be a necessary but not sufficient condition for an appropriate measurement invariance model. Consequently, while global fit measures indicate that loosening constraints on COM and CPV items is sufficient, concern for individual item DIF leads us to proceed to a third round of DIF testing.

In the third round, then, we constrain loadings for CPV3, CPV4, COM1, and COM4 to be equal across countries and iteratively test placing constraints on items in the SES and APS dimensions.

[Figure 5 around here]

Figure 5 shows the result of this analysis and confirms the expectation that APS1 and APS4 both show signs of DIF. Indeed, the estimated fit measure changes for these items exceed the similar estimates for COM and CPV items in previous analyses. Consequently, we loosen constraints on these two items as well.

In the resulting model, then, SES is estimated with constraints on all item loads, whereas APS is estimated with constraints only on APS2 and APS3, CPV is estimated with constraints only on CPV3 and CPV4, and COM is estimated with constraints only on COM1 and COM4. The resulting models fits the data well (χ^2 =3257.71, df = 847, p-value < 0.001; CFI = 0.979; RMSEA = 0.028) and shows fit deterioration well within the benchmarks ($\Delta\chi^2$ = 166.70; ΔCFI = 0.006; ΔRMSEA = 0.004).

We can only go through one additional round of DIF testing since only the SES dimension remains fully constrained. Doing so results in absolute fit measure changes indicating DIF in particularly SES4 and SES2 (not shown). Once again, the changes indicate substantial DIF comparable or even exceeding the changes in our first rounds. Consequently, in our final model, we constrain item loadings to be equal for APS2, APS3, CPV3, CPV4, COM1, COM4, SES1, and SES3 only, leaving half of the loadings unconstrained.

The resulting model fits the data well (χ^2 =3193.99, df = 840, p-value < 0.001; CFI = 0.980; RMSEA = 0.028) and shows fit deterioration comfortably within the benchmarks ($\Delta\chi^2$ = 102.99; ΔCFI = 0.005; ΔRMSEA = 0.003). Thus, we were able to construct a partially invariant measurement model that meets criteria for fit deterioration on global indices and, as best as possible, addresses DIF in individual items.

Turning for the first time to our validation data, we estimate a baseline model letting all factor loadings be freely estimated. Subsequently, we estimate our partial metric invariance model constraining all factor loadings but APS2, APS3, CPV3, CPV4, COM1, COM4, SES1, and SES3 to be equal across countries. Both the baseline (χ^2 =3012.20, df = 784, p-value < 0.001; CFI = 0.984; RMSEA = 0.026) and the partial metric

invariance models (χ^2 =3530.70, df = 840, p-value < 0.001; CFI = 0.976; RMSEA = 0.032) fit the data well. The fit deterioration from the former to the latter is within conventional benchmarks ($\Delta\chi^2$ = 518.50; ΔCFI = 0.009; ΔRMSEA = 0.006). In other words, our first-order partial metric invariance model validates on our validation data (see Appendix E).

6.3 Second-order metric invariance

Finding first-order partial metric invariance, we proceed to assess second-order cross-country metric invariance. This is a first in the PSM literature.[11] As noted above, we do so for a reflective second-order model.

The introduction of the reflective second-order construct slightly deteriorates fit for our partially metric invariant first-order model even when second-order factor loadings are estimated freely between groups. The models does, however, still fit the data well (χ^2 = 4316.80, df = 1.090, p-value < 0.001; CFI = 0.970; RMSEA = 0.033).

When testing full metric second-order invariance, we are forced to reject invariance as deterioration in global fit indices exceed our benchmark ($\Delta\chi^2$ = 453.27; ΔCFI = 0.016; ΔRMSEA = 0.007).[12] Consequently, we perform DIF testing for the second-order factor loadings, freeing all four second-order loadings and constraining one at a time.

Figure 6 shows the result of this analysis (see Appendix D for further details). As the figure indicates, the best fit is obtained by letting CPV and APS second-order factor-loadings vary across countries, leaving the required two second-order factor loadings – for SES and COM – fixed across countries.

[Figure 6 around here]

The resulting model not only fits the data well (χ^2 = 4275.75, df = 1.108, p-value < 0.001; CFI = 0.964; RMSEA = 0.035) but also falls below the deterioration benchmark for rejecting partial measurement equivalence ($\Delta\chi^2$ = 41.052; ΔCFI = 0.005; ΔRMSEA = 0.003). Consequently, for the second-order reflective model, we are able to establish partial second-order metric invariance in our training data. The caveat in

figure 6 is clear: while SES and, to a lesser extent, COM are relatively invariant in terms of loadings across countries, CPV and APS relate differently both to half or their items and to the PSM construct across countries.

Turning again to our validation data, we are once again able to validate our partial metric equivalence model. The fit deterioration between a model with unrestricted second-order factor loadings ($\chi^2 = 3680.53$, df = 870, p-value < 0.001; CFI = 0.971; RMSEA = 0.034) and a model restricting SES and COM to equality across countries ($\chi^2 = 3762.19$, df = 856, p-value < 0.001; CFI = 0.966; RMSEA = 0.037) is well within our benchmarks ($\Delta\chi^2 = 81.66$; ΔCFI = 0.005; ΔRMSEA = 0.003). Hence, we cannot reject second-order metric invariance on our validation data (see Appendix E). Our second-order reflective model is validated.

6.4 Scalar invariance

Using our partially metrically invariant model as a starting point, we next constrain item intercepts to be equal across countries. The fit of the resulting model is not impressive (χ^2 = 8359.48, df = 1.216, p-value < 0.001; CFI = 0.915; RMSEA = 0.052) and certainly worse than the metric invariance model ($\Delta\chi^2$ = 4083.73; ΔCFI = 0.049; ΔRMSEA = 0.017).

As a consequence, we next examine partial scalar invariance. Similar to our partial metric invariance test, we proceed by loosening all item intercepts and constraining one intercept reiteratively to determine DIF for each item. Also similar to our previous test, each dimension requires at least two items to be loosened, as effects coding identifies latent means by fixing the sum of item intercepts to zero. At least two item intercepts are required to be invariant for each dimension for the PSM construct to be first-order scalar invariant.

Our analysis failed to identify a partially scalar invariant model. Even fixing half of all item intercepts, fit deterioration from a model with freely estimated intercepts exceeds invariance benchmarks (see Appendix E for detailed results). Hence, PSM does not feature scalar invariance even in our sample of eight countries.

7 Discussion

Our analyses validated models supporting configurational invariance, as well as first- and second-order partial metric invariance for a reflective PSM construct in eight out of ten countries. Our two Asian cases were the sole exception. At the same time, our data did not support full or partial scalar invariance.

What does this mean for applied PSM research? Two answers. The first answer is positive: our findings imply that, contrary to the conclusion in Kim et al. (2013), our data supports some optimism that structural regression estimates are comparable across even very different countries using rigorous benchmarks for model evaluation. This is good news, for several reasons.

First, as PSM research continues to go far beyond the Anglo-American origins of the concepts and its measures, research can accumulate. Without metric invariance, comparative public management (Fitzpatrick et al. 2011) becomes difficult as we can only answer comparative questions qualitatively. With metric invariance, findings can be quantitatively compared. That is, our findings support concluding that the effect

of PSM on turnover intention is smaller or larger in, say, Ghana than in Brazil. This also implies that the findings of systematic literature reviews and meta analyses of PSM are meaningful, rather than invalid (e.g. Harari et al. 2016; Awan, Bel, and Esteve 2018; Ritz, Brewer, and Neumann 2016). With partial metric equivalence, the signs and size of coefficients of the causes and consequences of PSM can be compared across (most) countries.

Second, our findings validate the battery developed by Kim et al. (2013) in eight governments, excluding Nepal and Bangladesh. Through our cognitive interviews with public servants prior to fielding, we were able to find local language translations of PSM items which respondents across countries understood in a qualitatively comparable manner. In the collected survey data, the four-factor PSM construct fits well. We believe that, with Kim et al.'s (2013) work, PSM researchers have a solid measurement tool. If cross-national comparisons are to be valid, however, some adjustment may still be needed in South Asian cases, even if the construct displays acceptable but not good fit in those cases in our data.

A second answer from our data is negative: we were unable to establish full metric or (any) scalar invariance. Again, there are multiple consequences. First, scalar non-invariance implicates that means of PSM and of its dimensions are not comparable across countries. As comparisons of means of PSM and of PSM dimensions are not meaningful, cross-country PSM surveys cannot provide insights into which countries' public officials are more or less motivated to serve the public. This, unfortunately, both precludes PSM benchmarking between countries, and sheds doubt on the validity of PSM studies which derive their inferences from comparing PSM levels across countries (e.g. Vandenabeele and Van de Walle 2008). This conclusion is not due, moreover, to the rigorous benchmarks we use for model comparison. Recommendations for more lenient benchmarks in settings with many groups extend to metric invariance testing only, while standard benchmarks should be used for scalar invariance testing (Rutkowski and Svetina 2014). Hence, even if we were to use lenient model comparison benchmarks for our ten countries - which we argue is not appropriate - the conclusion would still include bad news for cross-national comparisons of PSM means.

Moreover, we established second-order partial metric invariance only through freely estimating 10 of 20 factor loadings. Self-sacrifice and, to a lesser extent, compassion were relatively invariant in terms of loadings across countries. At the same time, commitment to public values and attraction to public service relate differently both to half or their items and to the PSM construct across countries. From the perspective

of PSM as a type of motivation founded in public service, it is perhaps worrying that the two "most public" PSM dimensions appear to be the most culturally affected ones in terms of their measurement. This finding is not counter-intuitive. Public values may be different in different settings, leading to different associations and different common variance components of items related to public values across the globe.

Strictly speaking, however, we cannot be certain that the construct is in fact *culturally* affected. In principle, selection into public service could matter as well. Individuals with high PSM are often expected to seek careers in the public sector. However, as studies of dishonesty across national settings indicate, individuals with different types of characteristics select into public service in different contexts (Barfort et al. 2019; Hanna and Wang 2017). This may lead to differences in levels of PSM across countries but also - which is more relevant for our purposes - potentially to "public" PSM dimensions displaying the differences in structure we observe.

8 Conclusion

Based on a measurement invariance analysis of a 16-item PSM scale administered to 23,000 public servants in ten countries and four world regions – the, by far, largest original PSM survey in the literature to-date – our paper provides an empirical foundation for claims to a cross-cultural basis of PSM and cross-country knowledge accumulation in PSM research. At the same time, it underscores the limits of these claims, particularly when it comes to comparing PSM means across countries, applying PSM scales indiscriminately in Asia, and treating Commitment to Public Values and Attraction to Public Service as cross-country invariant PSM dimensions.

Beyond providing foundational evidence for cross-country knowledge accumulation (and its limits) in PSM research, our paper's findings point to several important areas for future research.

First, while our results suggests that Kim et al.'s (2013) scale provides a solid cross-country measurement tool, they also underscore that some adjustment may still be needed in Asian cases, where we found acceptable, but less than good fit - and measurement non-invariance even if we include the cases on lenient fit indices benchmarks. Future measurement research, from this perspective, ought to strive to build

adjustments to the battery such that it fits better in Asian cultural contexts, albeit in area comparisons with other world regions so we do not lose fit in other contexts by adapting to Asian cases.

Second, our finding that CPV and APS are relatively more cross-country non-invariant puts a premium on research to understand why and how the two "most public" PSM dimensions are affected in terms of their measurement. While public value research is ongoing in Europe and North America, very little parallel research exists in other parts of the World. Taking public values research global, ideally in comparative studies, constitutes one important avenue for understanding why some PSM dimensions behave somewhat differently in different cultural settings. Comparative public values is a topic ripe for both substantive and measurement research.

One possible route forward in this research is to focus on macro-factors. Recent developments in multilevel structural equation models (e.g. Davidov et al. 2012; Davidov et al. 2018) permit testing empirically which macro-level characteristics of nations give rise to differences in factor loadings and item intercepts. The obvious drawback of this strategy, of course, is that it requires collaborative projects on an unprecedented scale in order to have a sufficient number of nations represented for multilevel models to give adequate estimates, while being complex enough to identify the correct macro-level determinants of invariance. Multilevel tools for measurement invariance testing are an active area of research, and new options may become available. Until then, utilizing them to get answers related to full-scale, multidimensional PSM batteries requires a lot of shoe leather.

We believe these findings and implications add importantly to the literature on PSM and to comparative public management more generally, which remains characterized by a dearth of cross-country measurement equivalence analyses of survey scales. Our study suggests the feasibility of undertaking such analyses based on large-scale original cross-country survey data collection, and introduces to public administration measurement standards from cross-cultural survey research – in particular partial metric invariance – which can be used to robustly assess cross-country measurement equivalence of survey scales. At the same time, our study is, of course, not without limitations. Two stand out.

First, while the size of our sample enhances faith in the generalizability of our findings, it is nonetheless limited in three ways. First, it is tilted towards the developing world, comprising only two OECD countries (Estonia and Chile). We did not find partial metric non-invariance between the developing and OECD

countries in our sample, thus giving us no empirical reason to believe we would do so if other OECD countries – particularly in Weste1rn Europe and North America – were added to the sample. It remains for future research to more conclusively assess whether this is, in fact, the case, however. Second, our Asian cases (Bangladesh and Nepal) are distinct from the Asian cases that PSM research has largely focused on, in particular South Korea (e.g. Kim 2011), China (e.g. Liu and Perry 2016), and Taiwan (e.g. Chen, Hsieh, and Chen 2014). Whether the Asian 'exceptionalism' we see in our data also travels to these other Asian countries, equally remains for future cross-regional studies to assess. Third, while our survey samples appear to be representative on at least some demographics, national representativeness is as much a concern to our study as it is to other PSM research. It remains a challenge for future research to conduct more nationally representative PSM research in governments without sacrificing diversity of context.

Second, we assessed measurement invariance with a second-order reflective model – rather than the first-order reflective, second-order formative model of the construct recommended by Kim (2011). Estimating such a model involves the challenge of finding theoretical correlates of PSM, measured using multi-item batteries that are themselves invariant. Our data does not contain such batteries, and provided how common cross-national non-invariance is, finding candidates may be difficult in itself.[13] We leave it as a challenge for future research to test measurement invariance of PSM across cultures with formative models.

Notes

[1] For debates on the four-factor structure, see e.g. Perry 1996; Kim et al. 2013; Coursey and Pandey 2007.

[2] Neither the MGCFA framework nor partial invariance testing are the only possible options for our analysis. Instead of partial measurement invariance, recent developments in Bayesian structural equation modelling permit approximate measurement invariance testing, essentially abandoning the requirement that group differences in loadings and intercepts are either large enough to be a concern or exactly zero (e.g. Van De Schoot et al. 2013). Instead of the MGCFA framework, measurement invariance has been approached using IRT (e.g. Reise, Widaman, and Pugh 1993) or multilevel SEM (e.g. Davidov et al. 2012). Our choice of MGCFA is partly necessary – as we do not have enough groups for multilevel SEM estimates to be correct – and partly conventional as PSM researchers rely on CFA and SEM for their analyses rather than IRT.

[3] Kim et al. (2013, fn 11) use a different threshold since their analysis relies on LISREL, which calculates the CFI differently than most other software.

[4] Performing the analysis using the more lenient thresholds, as the reader can confirm from the following, results in the conclusion that full metric invariance obtains outside Asia. Scalar invariance, as it uses the same benchmarks regardless of the number of groups, does not. However, as noted in the main text, we consider the standard benchmarks more appropriate.

[5] We discarded two alternative approaches due to their limitations. A first alternative is to select items on conceptual grounds – that is, to determine theoretically which items loadings or intercepts are most likely to vary in different national settings. This comes with some obvious caveats as it introduces researcher discretion and interpretation into model building, with concomitant disagreements about the appropriateness of models and consequently results. A second alternative is data-driven and uses modification indices to determine which equality constraints give the largest reduction in model fit and proceed from that information. However, it is impossible for the researcher to know which of the recommended changes are sample specific. Consequently, any data driven approach to partial invariance risks building a model that cannot be replicated outside the sample used to build it (Putnick and Bornstein 2016).

[6] DIF is a term borrowed from item response theory. See Lee et al. (2018) for a discussion on the parallels between IRT and SEM, in particular MGCFA.

[7] Unfortunately, no benchmarks are available for changes in global fit indices when used for testing factorial invariance at the item level (Lee et al. (2018), 78).

[8] Our sample from Kosovo additionally covers some municipal employees.

[9] Kim et al.'s (2013) focus on a best fitting model is motivated by previous debates concerning the factor structure of PSM and the discriminant validity of the concepts' dimensions. As we are instead interested in the invariance of the four-factor PSM construct across national contexts, our benchmark for configurational invariance is simpler than Kim et al.'s (2013).

[10]The seemingly perfect fit for Kosovo and Malawi on the RMSEA and CFI is due to the χ^2 being smaller than the degrees of freedom.

[11]Kim et al. (2013) only focus on the dimensionality of the first order. Given that they do not find evidence of first-order (full) metric invariance, testing second-order metric invariance would have been superfluous, as establishment of the former is recommended before testing the latter.

[12]The model includes a Heywood case – for the variance of APS in Uganda. However, as the estimate is not significantly different from zero, we do not consider them evidence of misspecification (see Kolenikov and Bollen 2012).

[13]Building measurement models of formative constructs is not simple as these models, on their own, are not identified (Bollen and Lennox 1991). Three solutions to this problem are to: (1) include a reflective portion in the measurement model to identify it (Diamantopoulos and Papadopoulos 2010); (2) include endogenous manifest or latent variables affected by the formative construct in the model (forming a MIMIC model, as proposed for PSM by Kim 2011); or (3) identifying PSM as a composite. The first strategy involves changing the formative construct by including a reflective component in it. Diamantopoulos and Papadopoulos (2010, 363-364) propose a procedure in which metric invariance is established for the reflective portion of the construct prior to the formative portions being included. In their application, items are chosen for the reflective portion of the construct that "capture overall evaluations" (2010, 365) of the construct. Conceptually, this seems at odds with the purpose of having a formative measurement model in the first place: that each dimension of the construct is a separate component of it. For PSM, it is unclear which items should be chosen to reflect all aspects of the construct. Consequently, we do not rely on this strategy. The second strategy, some researchers have argued (Franke, Preacher, and Rigdon 2008; Howell, Breivik, and Wilcox 2007), may make the estimates of the effects of formative indicators on their construct sensitive to which variables are included as consequents of the latent variable. In the literature, this effect is sometimes referred to as interpretational confounding. However, as Bollen (2007) points out, such effects are due to structural misspecification and not to inherent sensitivity of the formative construct to its consequents. In other words, the choice of effect indicators or constructs does not introduce interpretational confounding in correctly specified models. Diamantopoulos and Papadopoulos (2010, 363) note that it is important to determine metric invariance for outcome scales before estimating effects of causal indicators on their latent, formative construct. Unfortunately, we do not have two other scales in our survey that fulfilled this requirement, and where model fit was sufficiently good for us to not suspect structural misspecification. From a measurement invariance perspective, the third strategy – constructing PSM as a composite of its dimensions – is not insightful. This strategy assumes what measurement invariance testing sets out to test, as slopes from dimensions to construct are identical across countries by design. While the literature does include models that allow weights on composites to be estimated freely rather than being fixed by the researcher (as applied PSM composites uniformly are), methodologists warn against the use of these strategies (e.g. Howell 2013; Lee, Cadogan, and Chamberlain 2013). We thus cannot assess a first-order reflective, second-order formative model of the PSM construct.

References

Awan, Sahar, Germa Bel, and Marc Esteve. 2018. "The benefits of PSM: an oasis or a mirage?" *IREA–Working Papers, 2018, IR18/25*.

Barfort, Sebastian, Nikolaj A Harmon, Frederik Hjorth, and Asmus Leth Olsen. 2019. "Sustaining honesty in public service: The role of selection". *American Economic Journal: Economic Policy* 11 (4): 96–123.

Behn, Robert D. 1995. "The Big Questions of Public Management". *Public administration review* 55 (4): 313–324.

Bollen, Kenneth A. 2007. "Interpretational confounding is due to misspecification, not to type of indicator: Comment on Howell, Breivik, and Wilcox (2007)." *Psychological methods* 12 (2): 219–228.

Bollen, Kenneth, and Richard Lennox. 1991. "Conventional wisdom on measurement: A structural equation perspective." *Psychological bulletin* 110 (2): 305.

Byrne, Barbara M. 2008. "Testing for multigroup equivalence of a measuring instrument: A walk through the process". *Psicothema* 20 (4): 872–882.

Chen, Chung-An, Chih-Wei Hsieh, and Don-Yun Chen. 2014. "Fostering public service motivation through workplace trust: Evidence from public managers in Taiwan". *Public Administration* 92 (4): 954–973.

Chen, Fang Fang, Karen H Sousa, and Stephen G West. 2005. "Testing measurement invariance of second-order factor models". *Structural equation modeling* 12 (3): 471–492.

Cheung, Gordon W, and Roger B Rensvold. 2002. "Evaluating goodness-of-fit indexes for testing measurement invariance". *Structural equation modeling* 9 (2): 233–255.

Christensen, Robert K, Laurie Paarlberg, and James L Perry. 2017. "Public service motivation research: Lessons for practice". *Public Administration Review* 77 (4): 529–542.

Clerkin, Richard M, and Jerrell D Coggburn. 2012. "The dimensions of public service motivation and sector work preferences". *Review of Public Personnel Administration* 32 (3): 209–235.

Coursey, David H, and Sanjay K Pandey. 2007. “Public service motivation measurement: Testing an abridged version of Perry’s proposed scale”. *Administration & Society* 39 (5): 547–568.

Davidov, Eldad, Hermann Dülmer, Jan Cieciuch, Anabel Kuntz, Daniel Seddig, and Peter Schmidt. 2018. “Explaining measurement nonequivalence using multilevel structural equation modeling: The case of attitudes toward citizenship rights”. *Sociological Methods & Research* 47 (4): 729–760.

Davidov, Eldad, Hermann Dülmer, Elmar Schlüter, Peter Schmidt, and Bart Meuleman. 2012. “Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance”. *Journal of Cross-Cultural Psychology* 43 (4): 558–575.

Davidov, Eldad, Peter Schmidt, and Jaak Billiet. 2018. *Cross-Cultural Analysis*. 2nd edition. Cheltenham: Routledge.

Diamantopoulos, Adamantios, and Nicolas Papadopoulos. 2010. “Assessing the cross-national invariance of formative measures: Guidelines for international business researchers”. *Journal of International Business Studies* 41 (2): 360–370.

Esteve, Marc, and Christian Schuster. 2019. *Motivating public employees*. Cambridge: Cambridge University Press.

Fitzpatrick, Jody, Malcolm Goggin, Tanya Heikkila, Donald Klingner, Jason Machado, and Christine Martell. 2011. “A new look at comparative public administration: Trends in research and an agenda for the future”. *Public Administration Review* 71 (6): 821–830.

Franke, George R, Kristopher J Preacher, and Edward E Rigdon. 2008. “Proportional structural effects of formative indicators”. *Journal of Business Research* 61 (12): 1229–1237.

Hanna, Rema, and Shing-Yi Wang. 2017. “Dishonesty and selection into public service: Evidence from India”. *American Economic Journal: Economic Policy* 9 (3): 262–90.

Harari, Michael B, David EL Herst, Heather R Parola, and Bruce P Carmona. 2016. “Organizational correlates of public service motivation: A meta-analysis of two decades of empirical research”. *Journal of Public Administration Research and Theory* 27 (1): 68–84.

Howell, Roy D. 2013. “Conceptual clarity in measurement—Constructs, composites, and causes: a commentary on Lee, Cadogan and Chamberlain”. *AMS review* 3 (1): 18–23.

Howell, Roy D, Einar Breivik, and James B Wilcox. 2007. “Reconsidering formative measurement.” *Psychological methods* 12 (2): 205.

Hu, Li-tze, and Peter M Bentler. 1999. “Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives”. *Structural equation modeling: a multidisciplinary journal* 6 (1): 1–55.

Kim, Sangmook. 2011. “Testing a revised measure of public service motivation: Reflective versus formative specification”. *Journal of Public Administration Research and Theory* 21 (3): 521–546.

Kim, Sangmook, Wouter Vandenabeele, Bradley E Wright, Lotte Bøgh Andersen, Francesco Paolo Cerase, Robert K Christensen, Céline Desmarais, Maria Koumenta, Peter Leisink, Bangcheng Liu, et al. 2013. “Investigating the structure and meaning of public service motivation across populations: Developing an international instrument and addressing issues of measurement invariance”. *Journal of Public Administration Research and Theory* 23 (1): 79–102.

Kolenikov, Stanislav, and Kenneth A Bollen. 2012. “Testing negative error variances: Is a Heywood case a symptom of misspecification?” *Sociological Methods & Research* 41 (1): 124–167.

Lee, Jaehoon, Todd D Little, and Kristopher J Preacher. 2018. “Methodological issues in using structural equation models for testing differential item functioning”. In *Cross-cultural analysis: Methods and applications*, edited by Eldad Davidov, Peter Schmidt, Jaak Billiet, and Bart Meuleman, 65–94. New York: Routledge.

Lee, Nick, John W Cadogan, and Laura Chamberlain. 2013. “The MIMIC model and formative variables: problems and solutions”. *AMS review* 3 (1): 3–17.

Little, Todd D, David W Slegers, and Noel A Card. 2006. “A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models”. *Structural Equation Modeling* 13 (1): 59–72.

Liu, Bangcheng, and James L Perry. 2016. “The psychological mechanisms of public service motivation: A two-wave examination”. *Review of Public Personnel Administration* 36 (1): 4–30.

Meyer-Sahling, Jan-Hinrik, and Kim Sass Mikkelsen. 2016. “Civil service laws, merit, politicization, and corruption: The perspective of public officials from five East European countries”. *Public administration* 94 (4): 1105–1123.

Meyer-Sahling, Jan-Hinrik, Kim Sass Mikkelsen, and Christian Schuster. 2017. “The Causal Effect of Public Service Motivation on Ethical Behavior in the Public Sector: Evidence from a Large-Scale Survey Experiment”. *Journal of Public Administration Research and Theory*.

Oliveros, Virginia, and Christian Schuster. 2018. “Merit, tenure, and bureaucratic behavior: Evidence from a conjoint experiment in the Dominican Republic”. *Comparative Political Studies* 51 (6): 759–792.

Perry, James L. 1996. “Measuring public service motivation: An assessment of construct reliability and validity”. *Journal of public administration research and theory* 6 (1): 5–22.

Perry, James L, and Annie Hondeghem. 2008. *Motivation in public management: The call of public service*. Oxford: Oxford University Press.

Prebble, Mark. 2016. “Has the study of public service motivation addressed the issues that motivated the study?” *The American Review of Public Administration* 46 (3): 267–291.

Putnick, Diane L, and Marc H Bornstein. 2016. “Measurement invariance conventions and reporting: The state of the art and future directions for psychological research”. *Developmental Review* 41:71–90.

Rainey, Hal G, and Paula Steinbauer. 1999. “Galloping elephants: Developing elements of a theory of effective government organizations”. *Journal of public administration research and theory* 9 (1): 1–32.

Reise, Steven P, Keith F Widaman, and Robin H Pugh. 1993. “Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance.” *Psychological bulletin* 114 (3): 552.

Ritz, Adrian, Gene A Brewer, and Oliver Neumann. 2016. “Public service motivation: A systematic literature review and outlook”. *Public Administration Review* 76 (3): 414–426.

Rosseel, Yves. 2012. “Lavaan: An R Package for Structural Equation Modeling”. *Journal of Statistical Software* 48 (2): 1–36.

Rutkowski, Leslie, and Dubravka Svetina. 2014. “Assessing the hypothesis of measurement invariance in the context of large-scale international surveys”. *Educational and Psychological Measurement* 74 (1): 31–57.

Van De Schoot, Rens, Anouck Kluytmans, Lars Tummers, Peter Lugtig, Joop Hox, and Bengt Muthén. 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.

Vandenabeele, Wouter, and Steven Van de Walle. 2008. “International differences in public service motivation: Comparing regions across the world”. In *Motivation in public management: The call of public service*, edited by James L Perry and Annie Hondeghem, 223–244. Oxford: Oxford University Press.

Yuan, Ke-Hai, and Peter M Bentler. 2004. “On chi-square difference and z tests in mean and covariance structure analysis when the base model is misspecified”. *Educational and Psychological measurement* 64 (5): 737–757.