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

Forthcoming in the International Public Management Journal

Kim Sass Mikkelsen*
Department of Social Sciences and Business
Roskilde University
Universitetsvej 1, 25.2, DK-4000 Roskilde
Denmark
Phone: +4546742801
Email: ksass@ruc.dk

Christian Schuster
School of Public Policy
University College London
29-31 Tavistock Square, London, WC1H 9QU
United Kingdom
Phone: +44 (0)203 108 6935
Email: c.schuster@ucl.ac.uk

Jan-Hinrik Meyer-Sahling
School of Politics and International Relation
University of Nottingham
University Park, Nottingham, NG7 2RD
United Kingdom
Phone: +44 (0) 115 84 67513
Email: j.meyer-sahling@nottingham.ac.uk

*corresponding author

Acknowledgement: Authors are ordered using a rotation principle. All authors contributed equally to this article. We thank Lotte Bøgh Andersen, Mogens Jin Pedersen, the participants at the 2019 Behavioral Public Administration Working Group at the 2019 European Group for Public Administration (EGPA), and two anonymous reviewers for their helpful comments on previous iterations of the manuscript. We gratefully acknowledge financial support for this research from the British Academy–UK Department for International Development (BA/DFID) Anti-Corruption Evidence Partnership.

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

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 ($\chi2$ = 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 (χ^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 (χ^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.

A Cross-Cultural Basis for Public Service? Appendices

Contents

A Country Characteristics

Table A1: Country Characteristics

	Income (World Bank 2018)	Democracy (Freedom House 2018)	Corruption Perception Score (Transparency International 2018, 100 indicates "very clean")
Albania	Upper middle income	Partly Free (68/100)	36/100
Bangladesh	Lower middle income	Partly Free (45/100)	26/100
Brazil	Upper middle income	Free (78/100)	35/100
Chile	High income	Free (94/100)	67/100
Estonia	High income	Free (94/100)	73/100
Ghana	Lower middle income	Free (83/100)	41/100
Kosovo	Lower middle income	Partly Free (52/100)	37/100
Malawi	Low income	Partly Free (63/100)	32/100
Nepal	Low income	Partly Free (55/100)	31/100
Uganda	Low income	Partly Free (37/100)	26/100

Sources: World Bank, 2018; Freedom House 2019; Transparency International, 2019.

B Survey Sample Descriptive Statistics and Representativeness

Table B1: Representativeness data for Chile

	Survey sample	Survey population[1] (all central government institutions)
Percentage female	56.1%	58%
Percentage university educated	49.6%	50%
Mean age	39	42[2]

Sources: Direccion de Presupuestos del Ministerio de Hacienda, 2017.
Notes: (1) Survey population data is from 2017; our survey was conducted in 2016-2017. (2) Estimated based on averaging age bands published in Direccion de Presupuestos del Ministerio de Hacienda de Chile (2019).

Table B2: Representativeness data for Ghana

	Survey sample (excl. administrative assistants)	Survey population data[1]
Percentage female	46.8%	45%
Percentage university educated	80.4%	70%
Mean age	35	42
Mean years of service	10.3	14

Sources: Rasul, Rogger, and Williams, 2015.
Notes: (1) Representative survey of staff in 45 Ministries and Departments in Accra, excluding public servants below technical-administrative grades. Survey population data is from 2015; our survey sample is from 2017.

Table B3: Representativeness data for Bangladesh

	Survey sample	Survey population data[1] (central government ministries and divisions)
Percentage female	22%	18%
Percentage managers	22%	27%

Source: Government of Bangladesh, 2018.
Notes: (1) Survey population data is from 2017; our survey was conducted in 2017.

Table B4: Representativeness data for Brazil

	Survey sample	Survey population data[1] (federal government)
Percentage female	43%	45%
Percentage university educated	90%	75%
Mean age	47.6	46

Source: Government of Brazil, 2018.
Notes: (1) Survey population data is from 2018; our survey was conducted in 2017.

Table B5: Representativeness data for Estonia

	Survey sample	Survey population data[1]
Percentage female	74%	56%
Mean age	43.6	43.3
Percentage university educated	70%	61%

Source: Government of Estonia, 2018.
Notes: (1) Population data for the Estonian civil service, including armed forces (which were excluded from our survey).

Table B6: Representativeness data for Uganda

	Survey sample	Survey population data[1]
Percentage female	45.5%	37.8%
Age: 24 or younger	0.7%	1.6%
Age: 25-29	9.1%	9.5%
Age: 30-34	25.3%	19.0%
Age: 35-39	21.8%	19.0%
Age: 40-44	18.0%	17.5%
Age: 45-49	11.3%	14.3%
Age: 50-54	7.1%	11.1%
Age: 55-59	4.2%	6.4%
Age: 60 and above	2.4%	1.6%

Source: Government of Uganda, 2019.
Notes: (1) 2019 data for the Ugandan public service as a whole (central and local government).

References

Direccion de Presupuestos del Ministerio de Hacienda. 2017. *Anuario Estadístico del Empleo Público en el Gobierno Central 2011-2018*. Santiago: Direccion de Presupuestos del Ministerio de Hacienda. URL: https://www.dipres.gob.cl/598/articles-191413_doc_pdf.pdf

Government of Bangladesh. 2018. *SUMMARY OF MANPOWER OF MINISTRIES AND DIVISIONS, DEPARTMENTS AND DIRECTORATES*. Dhaka: Government of Bangladesh. URL: http://old.mopa.gov.bd/uploads/2018/archive/manual/6582_(1-96)_part-2.pdf

Government of Brazil. 2018. *Painel Estatistico de Pessoal*. URL: http://painel.pep.planejamento.gov.br/QvAJAXZfc/opendoc.htm?document=painelpep.qvw&lang=en-US&host=Local&anonymous=true

Government of Estonia. 2018. *Personali- ja palgastatistika*. Tallinn: Government of Estonia. URL: https://www.rahandusministeerium.ee/et/riigi-personalipoliitika/personali-ja-palgastatistika

Government of Uganda. 2019. *Ministry of Public Service Newsletter June - Oct 2019*. Kampala: Ministry of Public Service. URL: https://publicservice.go.ug/download/ministry-of-public-service-newsletter-june-oct-2019/

Rasul, I, D. Rogger, and M. J. Williams. 2018. *Management and Bureaucratic Effectiveness: Evidence from the Ghanaian Civil Service*. Washington: The World Bank. URL: http://documents.worldbank.org/curated/en/335361537384686708/pdf/WPS8595.pdf.

Table B7: Descriptive Statistics I

Survey Item	Country	Mean	Std.Dev	Min	Max
APS1	Albania	3.735	0.563	0	4
APS2	Albania	3.740	0.568	0	4
APS3	Albania	3.792	0.525	0	4
APS4	Albania	3.804	0.485	0	4
COM1	Albania	3.511	0.770	0	4
COM2	Albania	3.579	0.706	0	4
COM3	Albania	3.704	0.747	0	4
COM4	Albania	3.368	0.890	0	4
CPV1	Albania	3.879	0.414	0	4
CPV2	Albania	3.875	0.377	0	4
CPV3	Albania	3.899	0.351	0	4
CPV4	Albania	3.949	0.298	0	4
SES1	Albania	3.170	0.864	0	4
SES2	Albania	3.072	1.059	0	4
SES3	Albania	2.618	1.124	0	4
SES4	Albania	3.313	0.878	0	4
APS1	Bangladesh	3.763	0.553	0	4
APS2	Bangladesh	3.632	0.605	0	4
APS3	Bangladesh	3.850	0.426	0	4
APS4	Bangladesh	3.644	0.633	0	4
COM1	Bangladesh	3.773	0.525	0	4
COM2	Bangladesh	3.607	0.663	0	4
COM3	Bangladesh	3.790	0.504	0	4
COM4	Bangladesh	3.742	0.574	0	4
CPV1	Bangladesh	3.597	0.871	0	4
CPV2	Bangladesh	3.358	0.938	0	4
CPV3	Bangladesh	3.595	0.793	0	4
CPV4	Bangladesh	3.909	0.341	0	4
SES1	Bangladesh	3.480	0.747	0	4
SES2	Bangladesh	3.281	0.972	0	4
SES3	Bangladesh	2.933	1.043	0	4
SES4	Bangladesh	3.245	0.898	0	4

Table B8: Descriptive Statistics II

Survey Item	Country	Mean	Std.Dev	Min	Max
APS1	Brazil	3.741	0.595	0	4
APS2	Brazil	3.713	0.641	0	4
APS3	Brazil	3.728	0.589	0	4
APS4	Brazil	3.807	0.509	0	4
COM1	Brazil	3.220	1.098	0	4
COM2	Brazil	3.281	0.987	0	4
COM3	Brazil	3.861	0.491	0	4
COM4	Brazil	3.706	0.603	0	4
CPV1	Brazil	3.854	0.465	0	4
CPV2	Brazil	3.844	0.512	0	4
CPV3	Brazil	3.841	0.439	0	4
CPV4	Brazil	3.978	0.213	0	4
SES1	Brazil	2.660	1.089	0	4
SES2	Brazil	3.013	1.067	0	4
SES3	Brazil	2.321	1.219	0	4
SES4	Brazil	2.751	1.172	0	4
APS1	Chile	3.776	0.524	0	4
APS2	Chile	3.762	0.540	0	4
APS3	Chile	3.872	0.422	0	4
APS4	Chile	3.845	0.444	0	4
COM1	Chile	3.469	0.805	0	4
COM2	Chile	3.687	0.576	0	4
COM3	Chile	3.801	0.529	0	4
COM4	Chile	3.647	0.631	0	4
CPV1	Chile	3.882	0.428	0	4
CPV2	Chile	3.851	0.457	0	4
CPV3	Chile	3.789	0.527	0	4
CPV4	Chile	3.937	0.347	0	4
SES1	Chile	3.057	0.941	0	4
SES2	Chile	2.995	1.017	0	4
SES3	Chile	2.562	1.201	0	4
SES4	Chile	3.097	1.025	0	4

Table B9: Descriptive Statistics III

Survey Item	Country	Mean	Std.Dev	Min	Max
APS1	Estonia	3.458	0.745	0	4
APS2	Estonia	3.070	0.834	0	4
APS3	Estonia	3.429	0.708	0	4
APS4	Estonia	3.144	0.766	0	4
COM1	Estonia	3.138	0.810	0	4
COM2	Estonia	3.176	0.807	0	4
COM3	Estonia	3.704	0.530	0	4
COM4	Estonia	3.288	0.724	0	4
CPV1	Estonia	3.519	0.692	0	4
CPV2	Estonia	3.712	0.532	0	4
CPV3	Estonia	3.650	0.604	0	4
CPV4	Estonia	3.812	0.472	0	4
SES1	Estonia	1.937	0.932	0	4
SES2	Estonia	1.739	0.954	0	4
SES3	Estonia	1.708	0.950	0	4
SES4	Estonia	2.324	0.915	0	4
APS1	Ghana	3.902	0.453	0	4
APS2	Ghana	3.855	0.493	0	4
APS3	Ghana	3.883	0.430	0	4
APS4	Ghana	3.904	0.401	0	4
COM1	Ghana	3.868	0.531	0	4
COM2	Ghana	3.812	0.566	0	4
COM3	Ghana	3.846	0.550	0	4
COM4	Ghana	3.873	0.493	0	4
CPV1	Ghana	3.900	0.464	0	4
CPV2	Ghana	3.661	0.812	0	4
CPV3	Ghana	3.873	0.549	0	4
CPV4	Ghana	3.865	0.598	0	4
SES1	Ghana	3.721	0.606	0	4
SES2	Ghana	3.565	0.773	0	4
SES3	Ghana	3.291	0.943	0	4
SES4	Ghana	3.659	0.672	0	4

Table B10: Descriptive Statistics IV

Survey Item	Country	Mean	Std.Dev	Min	Max
APS1	Kosovo	3.878	0.462	0	4
APS2	Kosovo	3.784	0.555	0	4
APS3	Kosovo	3.897	0.461	0	4
APS4	Kosovo	3.883	0.458	0	4
COM1	Kosovo	3.690	0.782	0	4
COM2	Kosovo	3.754	0.601	0	4
COM3	Kosovo	3.885	0.451	0	4
COM4	Kosovo	3.798	0.584	0	4
CPV1	Kosovo	3.897	0.414	1	4
CPV2	Kosovo	3.853	0.491	0	4
CPV3	Kosovo	3.802	0.609	0	4
CPV4	Kosovo	3.851	0.502	0	4
SES1	Kosovo	3.540	0.789	0	4
SES2	Kosovo	3.591	0.816	0	4
SES3	Kosovo	3.011	1.098	0	4
SES4	Kosovo	3.526	0.810	0	4
APS1	Malawi	3.684	0.662	0	4
APS2	Malawi	3.663	0.650	0	4
APS3	Malawi	3.689	0.668	0	4
APS4	Malawi	3.684	0.627	0	4
COM1	Malawi	3.653	0.730	0	4
COM2	Malawi	3.653	0.703	0	4
COM3	Malawi	3.682	0.675	0	4
COM4	Malawi	3.668	0.718	0	4
CPV1	Malawi	3.549	0.729	0	4
CPV2	Malawi	3.555	0.791	0	4
CPV3	Malawi	3.570	0.732	0	4
CPV4	Malawi	3.619	0.665	0	4
SES1	Malawi	3.612	0.762	0	4
SES2	Malawi	3.581	0.795	0	4
SES3	Malawi	3.521	0.837	0	4
SES4	Malawi	3.599	0.740	0	4

Table B11: Descriptive Statistics V

Survey Item	Country	Mean	Std.Dev	Min	Max
APS1	Nepal	3.876	0.386	1	4
APS2	Nepal	3.857	0.404	1	4
APS3	Nepal	3.883	0.437	0	4
APS4	Nepal	3.878	0.402	0	4
COM1	Nepal	3.886	0.367	1	4
COM2	Nepal	3.841	0.437	1	4
COM3	Nepal	3.849	0.472	0	4
COM4	Nepal	3.871	0.440	0	4
CPV1	Nepal	3.831	0.505	0	4
CPV2	Nepal	3.681	0.721	0	4
CPV3	Nepal	3.871	0.436	0	4
CPV4	Nepal	3.923	0.329	0	4
SES1	Nepal	3.649	0.645	0	4
SES2	Nepal	3.711	0.643	0	4
SES3	Nepal	2.898	1.153	0	4
SES4	Nepal	3.074	1.144	0	4
APS1	Uganda	3.811	0.587	0	4
APS2	Uganda	3.701	0.628	0	4
APS3	Uganda	3.760	0.582	0	4
APS4	Uganda	3.695	0.632	0	4
COM1	Uganda	3.672	0.700	0	4
COM2	Uganda	3.698	0.603	0	4
COM3	Uganda	3.714	0.678	0	4
COM4	Uganda	3.710	0.641	0	4
CPV1	Uganda	3.773	0.586	0	4
CPV2	Uganda	3.340	1.107	0	4
CPV3	Uganda	3.777	0.593	0	4
CPV4	Uganda	3.818	0.549	0	4
SES1	Uganda	3.132	1.085	0	4
SES2	Uganda	3.242	0.988	0	4
SES3	Uganda	2.654	1.261	0	4
SES4	Uganda	3.091	1.118	0	4

C Country Specific Model Results

Table C1: Factor Loadings for Country Specific Models

Item	Albania	Bangladesh	Brazil	Chile	Estonia
APS1	1.161 (0.071)	0.915 (0.090)	1.005 (0.050)	1.071 (0.029)	0.870 (0.036)
APS2	1.097 (0.056)	1.230 (0.116)	1.154 (0.053)	1.166 (0.034)	1.163 (0.036)
APS3	0.867 (0.076)	0.699 (0.125)	0.929 (0.051)	0.812 (0.035)	0.851 (0.035)
APS4	0.875 (0.047)	1.156 (0.105)	0.911 (0.049)	0.950 (0.024)	1.116 (0.029)
CPV1	1.429 (0.129)	0.913 (0.167)	1.378 (0.100)	1.002 (0.033)	1.362 (0.075)
CPV2	0.902 (0.121)	1.565 (0.223)	1.263 (0.087)	1.044 (0.041)	0.894 (0.050)
CPV3	0.946 (0.103)	0.912 (0.128)	0.988 (0.078)	1.302 (0.055)	1.059 (0.067)
CPV4	0.724 (0.112)	0.610 (0.173)	0.370 (0.127)	0.653 (0.055)	0.685 (0.058)
COM1	0.963 (0.068)	0.834 (0.099)	1.232 (0.062)	1.108 (0.033)	1.162 (0.032)
COM2	0.990 (0.042)	1.465 (0.177)	1.264 (0.058)	0.990 (0.026)	1.173 (0.030)
COM3	0.945 (0.066)	0.676 (0.137)	0.559 (0.065)	0.731 (0.036)	0.632 (0.038)
COM4	1.101 (0.067)	1.025 (0.111)	0.945 (0.053)	1.170 (0.030)	1.033 (0.036)
SES1	0.971 (0.055)	0.809 (0.075)	1.048 (0.025)	0.992 (0.023)	0.984 (0.029)
SES2	0.972 (0.048)	1.044 (0.092)	0.825 (0.036)	0.913 (0.024)	0.963 (0.030)
SES3	1.187 (0.039)	1.233 (0.101)	1.052 (0.028)	1.085 (0.025)	1.056 (0.028)
SES4	0.870 (0.042)	0.914 (0.093)	1.075 (0.029)	1.011 (0.026)	0.996 (0.037)
APS	1.010 (0.100)	0.958 (0.075)	1.114 (0.057)	0.978 (0.043)	1.349 (0.051)
CPV	0.341 (0.066)	0.898 (0.076)	0.487 (0.076)	0.639 (0.049)	0.654 (0.044)
COM	1.237 (0.097)	1.041 (0.068)	1.046 (0.059)	1.224 (0.046)	0.907 (0.044)
SES	1.412 (0.174)	1.103 (0.096)	1.353 (0.102)	1.159 (0.082)	1.090 (0.060)

Item	Ghana	Kosovo	Malawi	Nepal	Uganda
APS1	0.927 (0.041)	0.896 (0.080)	0.922 (0.060)	0.985 (0.093)	0.790 (0.064)
APS2	1.070 (0.048)	1.130 (0.107)	1.051 (0.054)	0.914 (0.066)	1.089 (0.046)
APS3	1.019 (0.028)	0.852 (0.066)	1.102 (0.056)	1.020 (0.096)	0.980 (0.051)
APS4	0.985 (0.027)	1.121 (0.086)	0.925 (0.091)	1.081 (0.092)	1.140 (0.050)
CPV1	0.896 (0.054)	0.704 (0.125)	1.111 (0.070)	0.632 (0.151)	0.978 (0.046)
CPV2	1.136 (0.078)	1.138 (0.087)	1.023 (0.082)	1.597 (0.208)	1.446 (0.106)
CPV3	1.023 (0.057)	1.106 (0.116)	0.966 (0.075)	1.142 (0.257)	0.857 (0.059)
CPV4	0.945 (0.049)	1.052 (0.128)	0.900 (0.051)	0.629 (0.116)	0.719 (0.064)
COM1	0.968 (0.050)	0.976 (0.113)	1.040 (0.069)	0.878 (0.093)	1.020 (0.030)
COM2	1.107 (0.058)	1.306 (0.111)	1.021 (0.044)	1.096 (0.085)	0.987 (0.038)
COM3	0.997 (0.035)	0.753 (0.116)	0.950 (0.072)	1.169 (0.083)	1.026 (0.029)
COM4	0.928 (0.049)	0.965 (0.089)	0.989 (0.052)	0.857 (0.098)	0.966 (0.035)
SES1	0.912 (0.043)	0.993 (0.053)	0.994 (0.049)	0.904 (0.082)	1.077 (0.036)
SES2	1.053 (0.037)	0.916 (0.059)	1.016 (0.051)	0.930 (0.089)	0.801 (0.053)
SES3	1.060 (0.050)	1.089 (0.075)	1.054 (0.063)	1.153 (0.080)	1.014 (0.041)
SES4	0.974 (0.048)	1.003 (0.052)	0.935 (0.060)	1.014 (0.086)	1.108 (0.034)
APS	0.993 (0.075)	0.881 (0.170)	0.839 (0.146)	1.179 (0.108)	0.852 (0.048)
CPV	1.032 (0.074)	0.591 (0.203)	1.105 (0.104)	0.670 (0.132)	0.953 (0.044)
COM	0.951 (0.082)	1.107 (0.143)	0.890 (0.107)	0.978 (0.108)	0.963 (0.041)
SES	1.024 (0.111)	1.421 (0.185)	1.166 (0.120)	1.173 (0.147)	1.232 (0.076)

Robust DWLS estimates for second-order CFA with standard errors in parentheses. All factor loadings are significant at a 1% α level.

Table C2: Fit Measures for Country Specific Models

country	CFI	χ^2	df	RMSEA
Albania	0.982	228.225	100	0.026
Bangladesh	0.919	306.257	100	0.037
Brazil	0.983	414.866	100	0.025
Chile	0.986	655.762	100	0.023
Estonia	0.959	1090.754	100	0.057
Ghana	0.994	173.806	100	0.010
Kosovo	1.000	77.431	100	0.000
Malawi	1.000	60.889	100	0.000
Nepal	0.920	290.536	100	0.037
Uganda	0.992	348.507	100	0.022

Robust CFI, scaled χ^2, and robust RMSEA fit indices by country. Note that χ^2 statistics are reported as χ^2 per respondent in the main text. All χ^2 tests are significant at a 1% α level.

Table C3: Mean and Standard Deviation of Factor Loadings from Country Specific Models across Countries

Item	Mean factor loading	Standard deviation
APS1	0.954	0.106
APS2	1.106	0.086
APS3	0.913	0.120
APS4	1.026	0.107
CPV1	1.041	0.278
CPV2	1.201	0.259
CPV3	1.030	0.129
CPV4	0.729	0.195
COM1	1.018	0.123
COM2	1.140	0.162
COM3	0.844	0.200
COM4	0.998	0.089
SES1	0.968	0.077
SES2	0.943	0.085
SES3	1.098	0.070
SES4	0.990	0.071
APS	1.015	0.160
CPV	0.737	0.249
COM	1.034	0.122
SES	1.213	0.139

Based on robust DWLS estimates for second-order CFA. Means are calculated as the average of the ten factor loadings from country specific models. Standard deviations are calculated as the standard deviation from those same ten numbers.

D DIF Testing for First Order Partial Metric Invariance

Table D1: First Round of Testing

Constrained item factor loading	CFI_{scalar}	$RMSEA_{scalar}$	χ^2_{scalar}	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$
APS1	0.984	0.026	3237.742	-0.00138	0.00103	146.741
APS2	0.986	0.024	3005.103	-0.00007	0.00016	85.897
APS3	0.985	0.025	3080.817	-0.00054	0.00034	10.183
APS4	0.984	0.025	3202.770	-0.00122	0.00090	111.769
CPV1	0.983	0.026	3334.484	-0.00251	0.00193	243.484
CPV2	0.983	0.026	3321.992	-0.00228	0.00175	230.992
CPV3	0.984	0.026	3186.532	-0.00131	0.00097	95.532
CPV4	0.984	0.026	3137.347	-0.00164	0.00124	46.346
COM1	0.985	0.025	3092.372	-0.00060	0.00039	1.372
COM2	0.984	0.026	3255.475	-0.00140	0.00104	164.475
COM3	0.982	0.027	3405.126	-0.00323	0.00248	314.125
COM4	0.985	0.025	3147.906	-0.00084	0.00059	56.905
SES1	0.985	0.025	3045.252	-0.00012	0.00001	45.749
SES2	0.985	0.025	3125.775	-0.00080	0.00056	34.774
SES3	0.985	0.025	3042.637	-0.00018	0.00004	48.364
SES4	0.985	0.025	3111.640	-0.00063	0.00041	20.639

Fit measures and differences in fit measures for model comparisons between the configurational model and a model constraining only the item in the first column to be equal across countries. Note ΔCFI is reported as absolute change in the main text. The models omit the second-order construct.

Table D2: Second Round of Testing

Constrained item factor loading	CFI_{scalar}	$RMSEA_{scalar}$	χ^2_{scalar}	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$
APS1	0.982	0.027	3321.621	-0.00139	0.00097	139.084
APS2	0.984	0.026	3104.318	-0.00007	0.00017	78.220
APS3	0.983	0.026	3175.137	-0.00054	0.00031	7.400
APS4	0.983	0.027	3288.345	-0.00122	0.00083	105.808
CPV1	0.981	0.028	3414.546	-0.00254	0.00182	232.009
CPV2	0.982	0.027	3399.951	-0.00229	0.00164	217.413
CPV3	0.983	0.027	3271.983	-0.00131	0.00090	89.445
CPV4	0.982	0.027	3225.860	-0.00164	0.00116	43.323
SES1	0.984	0.026	3142.631	-0.00013	0.00001	39.906
SES2	0.983	0.026	3215.704	-0.00080	0.00051	33.166
SES3	0.984	0.026	3138.980	-0.00017	0.00003	43.557
SES4	0.983	0.026	3201.444	-0.00062	0.00037	18.906

Fit measures and differences in fit measures for model comparisons between the configurational model and a model constraining COM1, COM4, and the item in the first column to be equal across countries. Note ΔCFI is reported as absolute change in the main text. The models omit the second-order construct.

Table D3: Third Round of Testing

Constrained item factor loading	CFI_{scalar}	$RMSEA_{scalar}$	χ^2_{scalar}	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$
APS1	0.979	0.029	3457.506	-0.00454	0.00301	274.968
APS2	0.981	0.028	3264.038	-0.00309	0.00199	81.500
APS3	0.980	0.028	3326.637	-0.00369	0.00242	144.099
APS4	0.979	0.029	3427.833	-0.00436	0.00289	245.296
SES1	0.981	0.028	3300.604	-0.00329	0.00213	118.067
SES2	0.980	0.028	3358.047	-0.00390	0.00257	175.509
SES3	0.980	0.028	3297.194	-0.00334	0.00217	114.657
SES4	0.980	0.028	3347.736	-0.00374	0.00246	165.199

Fit measures and differences in fit measures for model comparisons between the configurational model and a model constraining COM1, COM4, CPV3, CPV4, and the item in the first column to be equal across countries. Note ΔCFI is reported as absolute change in the main text. The models omit the second-order construct.

Table D4: Fourth Round of Testing

Constrained item factor loading	CFI_{scalar}	$RMSEA_{scalar}$	χ^2_{scalar}	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$
SES1	0.977	0.030	3580.193	-0.00639	0.00402	397.656
SES2	0.977	0.030	3633.105	-0.00700	0.00442	450.568
SES3	0.977	0.030	3576.440	-0.00644	0.00405	393.903
SES4	0.977	0.030	3625.048	-0.00685	0.00432	442.511

Fit measures and differences in fit measures for model comparisons between the configurational model and a model constraining COM1, COM4, CPV3, CPV4, APS2, APS3, and the item in the first column to be equal across countries. Note ΔCFI is reported as absolute change in the main text. The models omit the second-order construct.

Table D5: DIF Testing for Second Order Partial Metric Invariance

Constrained dimension factor loading	CFI_{scalar}	$RMSEA_{scalar}$	χ^2_{scalar}	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$
APS	0.969	0.034	3754.612	-0.01458	0.00839	572.075
CPV	0.962	0.038	4068.002	-0.02153	0.01205	885.465
COM	0.972	0.033	3621.824	-0.01227	0.00708	439.287
SES	0.974	0.031	3155.721	-0.00935	0.00535	26.817

Fit measures and differences in fit measures for model comparisons between the configurational model and a model constraining COM1, COM4, CPV3, CPV4, APS2, APS3, SES1, SES3, and the dimension in the first column to be equal across countries. Note ΔCFI is reported as absolute change in the main text.

E Validation for Metric Invariance

Table E1: Comparison of Model Fit for Validation Data (First Order Metric Invariance)

	CFI	RMSEA	χ^2	df
Configurational Invariance Model	0.984	0.026	$3,012.202$	784
Partial Metric Invariance Model	0.976	0.032	$3,530.701$	840
Δ	-0.009	0.006	518.500	56

Fit measures and differences in fit measures for model comparisons between the configurational model and a model constraining COM1, COM4, CPV3, CPV4, APS2, APS3, SES1, SES3 to be equal across countries. The models omit the second-order construct.

Table E2: Comparison of Model Fit for Validation Data (Second Order Metric Invariance)

	CFI	RMSEA	χ^2	df
Configurational Invariance Model	0.971	0.034	$3,762.198$	856
Partial Metric Invariance Model	0.966	0.037	$3,680.533$	870
Δ	-0.005	0.003	81.664	14

Fit measures and differences in fit measures for model comparisons between the configurational model and a model constraining COM1, COM4, CPV3, CPV4, APS2, APS3, SES1, SES3, COM, and SES to be equal across countries.

Table E3: Factor Loadings for Second-Order Partial Metric Invariance Validation Model

Loading	Uganda	Ghana	Malawi	Brazil
APS1	0.804 (0.067)	0.993 (0.023)	0.858 (0.064)	1.034 (0.037)
APS2	1.158 (0.017)	1.158 (0.017)	1.158 (0.017)	1.158 (0.017)
APS3	0.860 (0.018)	0.860 (0.018)	0.860 (0.018)	0.860 (0.018)
APS4	1.179 (0.066)	0.990 (0.023)	1.125 (0.065)	0.949 (0.037)
CPV1	0.848 (0.062)	1.027 (0.044)	1.021 (0.050)	1.210 (0.083)
CPV2	1.335 (0.071)	1.156 (0.047)	1.162 (0.051)	0.974 (0.081)
CPV3	1.084 (0.025)	1.084 (0.025)	1.084 (0.025)	1.084 (0.025)
CPV4	0.733 (0.026)	0.733 (0.026)	0.733 (0.026)	0.733 (0.026)
COM1	1.140 (0.017)	1.140 (0.017)	1.140 (0.017)	1.140 (0.017)
COM2	0.858 (0.039)	0.927 (0.029)	0.784 (0.052)	1.214 (0.047)
COM3	0.961 (0.038)	0.892 (0.030)	1.034 (0.052)	0.605 (0.048)
COM4	1.041 (0.017)	1.041 (0.017)	1.041 (0.017)	1.041 (0.017)
SES1	1.048 (0.012)	1.048 (0.012)	1.048 (0.012)	1.048 (0.012)
SES2	0.869 (0.038)	1.019 (0.036)	0.962 (0.048)	0.845 (0.028)
SES3	1.032 (0.014)	1.032 (0.014)	1.032 (0.014)	1.032 (0.014)
SES4	1.051 (0.037)	0.900 (0.037)	0.958 (0.048)	1.075 (0.027)
APS	0.857 (0.040)	0.912 (0.037)	0.688 (0.125)	1.251 (0.050)
CPV	0.898 (0.041)	0.842 (0.037)	1.066 (0.123)	0.503 (0.048)
COM	1.053 (0.021)	1.053 (0.021)	1.053 (0.021)	1.053 (0.021)
SES	1.193 (0.031)	1.193 (0.031)	1.193 (0.031)	1.193 (0.031)

Loading	Albania	Kosovo	Estonia	Chile
APS1	1.066 (0.052)	1.068 (0.073)	0.870 (0.034)	0.980 (0.022)
APS2	1.158 (0.017)	1.158 (0.017)	1.158 (0.017)	1.158 (0.017)
APS3	0.860 (0.018)	0.860 (0.018)	0.860 (0.018)	0.860 (0.018)
APS4	0.917 (0.051)	0.914 (0.071)	1.113 (0.032)	1.003 (0.020)
CPV1	1.102 (0.053)	1.116 (0.103)	1.195 (0.066)	1.078 (0.029)
CPV2	1.081 (0.057)	1.067 (0.103)	0.988 (0.062)	1.105 (0.031)
CPV3	1.084 (0.025)	1.084 (0.025)	1.084 (0.025)	1.084 (0.025)
CPV4	0.733 (0.026)	0.733 (0.026)	0.733 (0.026)	0.733 (0.026)
COM1	1.140 (0.017)	1.140 (0.017)	1.140 (0.017)	1.140 (0.017)
COM2	1.132 (0.052)	0.921 (0.039)	1.236 (0.030)	1.026 (0.022)
COM3	0.687 (0.053)	0.898 (0.040)	0.583 (0.032)	0.793 (0.025)
COM4	1.041 (0.017)	1.041 (0.017)	1.041 (0.017)	1.041 (0.017)
SES1	1.048 (0.012)	1.048 (0.012)	1.048 (0.012)	1.048 (0.012)
SES2	1.085 (0.047)	0.939 (0.057)	0.941 (0.032)	0.880 (0.019)
SES3	1.032 (0.014)	1.032 (0.014)	1.032 (0.014)	1.032 (0.014)
SES4	0.835 (0.048)	0.981 (0.057)	0.979 (0.033)	1.040 (0.020)
APS	1.033 (0.097)	1.066 (0.071)	1.231 (0.040)	1.111 (0.031)
CPV	0.721 (0.091)	0.688 (0.074)	0.523 (0.034)	0.643 (0.036)
COM	1.053 (0.021)	1.053 (0.021)	1.053 (0.021)	1.053 (0.021)
SES	1.193 (0.031)	1.193 (0.031)	1.193 (0.031)	1.193 (0.031)

Factor loadings fitted on the validation dataset obtaining second order partial metric invariance by constraining COM1, COM4, CPV3, CPV4, APS2, APS3, SES1, SES3, COM, and SES to be equal across countries.

F DIF Testing for First Order Partial Scalar Invariance

Table F1: First Round of Testing

Constrained item intercept	CFI_{scalar}	$RMSEA_{scalar}$	χ^2_{scalar}	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$
APS1	0.967	0.035	3605.534	-0.00281	0.00138	201.826
APS2	0.969	0.034	3466.572	-0.00070	0.00025	62.865
APS3	0.968	0.034	3512.216	-0.00126	0.00055	108.508
APS4	0.967	0.035	3593.883	-0.00266	0.00131	190.175
CPV1	0.966	0.035	3637.031	-0.00340	0.00170	233.323
CPV2	0.966	0.036	3658.180	-0.00365	0.00183	254.472
CPV3	0.967	0.035	3645.154	-0.00289	0.00143	241.446
CPV4	0.968	0.035	3559.457	-0.00183	0.00086	155.749
COM1	0.965	0.036	3808.230	-0.00488	0.00247	404.522
COM2	0.966	0.036	3679.447	-0.00393	0.00198	275.739
COM3	0.965	0.036	3707.154	-0.00441	0.00222	303.446
COM4	0.965	0.036	3801.981	-0.00475	0.00240	398.273
SES1	0.968	0.034	3513.470	-0.00126	0.00055	109.762
SES2	0.966	0.036	3689.269	-0.00383	0.00192	285.561
SES3	0.966	0.036	3721.218	-0.00373	0.00187	317.510
SES4	0.967	0.035	3566.876	-0.00234	0.00114	163.169

Table F2: Second Round of Testing

Constrained item intercept	CFI_{scalar}	$RMSEA_{scalar}$	χ^2_{scalar}	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$
APS1	0.959	0.039	4144.432	-0.01020	0.00480	740.724
APS2	0.961	0.038	4017.137	-0.00813	0.00380	613.429
APS3	0.961	0.038	4061.177	-0.00868	0.00407	657.469
APS4	0.960	0.039	4132.767	-0.01005	0.00473	729.059
CPV1	0.959	0.039	4170.880	-0.01077	0.00507	767.173
CPV2	0.959	0.039	4190.941	-0.01102	0.00519	787.233
CPV3	0.959	0.039	4190.359	-0.01031	0.00485	786.651
CPV4	0.960	0.038	4106.704	-0.00924	0.00434	702.996
SES1	0.961	0.038	4061.875	-0.00867	0.00406	658.167
SES2	0.958	0.039	4228.666	-0.01125	0.00530	824.959
SES3	0.958	0.039	4264.985	-0.01114	0.00525	861.277
SES4	0.960	0.038	4108.413	-0.00976	0.00459	704.705

Table F3: Third Round of Testing

Constrained item intercept	CFI_{scalar}	$RMSEA_{scalar}$	χ^2_{scalar}	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$
APS1	0.956	0.040	4402.740	-0.01383	0.00618	999.032
APS2	0.958	0.039	4280.757	-0.01178	0.00524	877.049
APS3	0.957	0.039	4324.208	-0.01234	0.00549	920.500
APS4	0.956	0.040	4391.319	-0.01368	0.00611	987.611
CPV1	0.955	0.040	4429.146	-0.01443	0.00644	1025.438
CPV2	0.955	0.040	4448.981	-0.01468	0.00656	1045.273
CPV3	0.956	0.040	4451.833	-0.01397	0.00624	1048.125
CPV4	0.957	0.040	4368.506	-0.01289	0.00575	964.798

Table F4: Fourth Round of Testing

Constrained item intercept	CFI_{scalar}	$RMSEA_{scalar}$	χ^2_{scalar}	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$
APS1	0.948	0.043	5069.346	-0.02211	0.00942	1665.638
APS2	0.950	0.042	4955.297	-0.02007	0.00857	1551.589
APS3	0.949	0.043	4998.105	-0.02061	0.00880	1594.397
APS4	0.948	0.043	5057.716	-0.02195	0.00936	1654.008

Table F5: Fit indices for first-order partial scalar invariance testing

Model	ΔCFI	$\Delta RMSEA$	$\Delta\chi^2$	Unconstrained intercepts
Full Scalar	-0.03342	0.01297	2485.434	
DIF Round I	-0.02785	0.01105	2052.040	COM1, COM4
DIF Round II	-0.02355	0.00952	1715.089	COM1, COM4, SES2, SES3
DIF Round III	-0.02027	0.00834	1520.796	COM1, COM4, SES2, SES3, CPV1, CPV2
DIF Round IV	-0.01755	0.00735	1346.400	COM1, COM4, SES2, SES3, CPV1, CPV2, APS1, APS4

All fit measure changes are relative to the fitted and validated second-order metric invariance model.

G Second-Order Modelling Choices in the PSM literature

To see whether our second-order modelling choices reflect common practice in the PSM literature as much as possible, we reviewed second-order modelling choices in PSM studies. To narrow the scope of our investigation to a feasible volume, we guided our selection of contributions as follows. We selected published journal articles from public administration and management journals that estimated PSM as part of their statistical models. As we were primarily interested in modelling choices for multidimensional PSM constructs with thought given to second-order modelling strategy, we narrowed this pool further to contributions citing either Kim (2010) or Kim et al. (2013) as well as those two contributions themselves. Both these contributions are well-cited and clearly advocate a multidimensional conception of PSM. Furthermore, we use Kim et al.'s (2013) battery for our statistical tests.

We coded modelling choices into six groups. One group (Single dimension) treats, despite our search strategy, PSM as a one-dimensional construct, typically using a reduced 4-6 item battery. Another group (Composite) estimates dimensions of PSM and aggregates a PSM composite from the dimensions, typically either through summing or averaging factor scores. A third and fourth group (Reflective and Formative) estimate second-order CFA models treating PSM either as a cause of or as caused by its dimensions. A fifth group of contributions (None) estimated and analysed dimensions without forming a higher-order construct. Finally, a sixth group included a higher-order PSM construct, but it was unclear how that construct is modelled. Excluding the latter group, and including a few studies as two entries if multiple strategies were pursued (mostly single dimension and none), our search resulted in 100 models from 97 published studies, all published after 2010 (see Appendix B for a full list of studies).

Figure 1 shows the distribution of these models across the five groups. As the figure shows, a large minority (41%) of our review database entries did not analyse PSM as a multidimensional construct. Among the remaining entries, separate analysis of dimensions (27%) and forming composites (17%) were more popular than formative and reflective strategies combined. These comprised 15% of the entries, just two of which were formative measurement models.

From the vantage point of this sample of literature, measurement invariance at the first order would suffice for most applied research either because their research does not consider a second-order construct (None) or because they construct composites directly from dimensions (Composite). The latter strategy does have a second conceptual level, and researchers frequently refer to it as formative. But since it assumes equality of factor loadings across groups by design – and hence assumes what measurement invariance models set out to test – it cannot be subjected to measurement invariance testing.

Though quite a few researchers take Kim's (2010) argument that PSM ought to be a second-order formative construct to heart – clearly more may do so in our sample due to our selection criteria – no one in our sample uses the formative latent variable model he proposed (the two formative entries in the database are both by Kim 2010; 2012).[1] For measurement invariance, this means that the only modelling strategy including a testable second-order latent construct applied somewhat frequently (in 12% of entries) is reflective. Our paper thus focuses on testing a reflective model.

[1] Kim (2012) uses a partial least squares estimation technique, which models composites with error. Hence, our conclusions in the main text in fact understates the rarity of formative measurement models in the literature we reviewed.

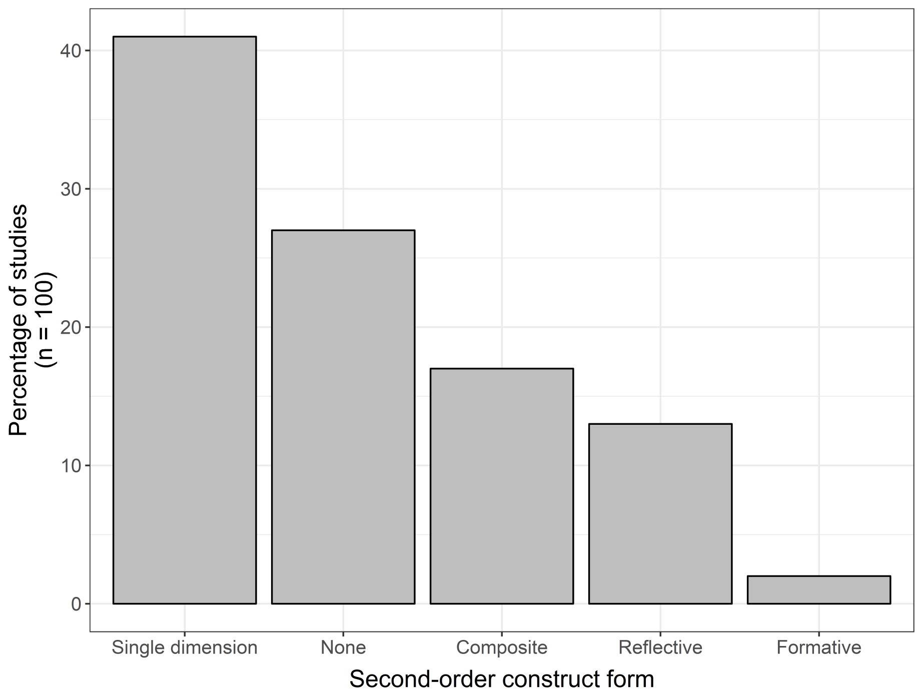

Figure 1: Review of second-order models in PSM research

H Articles Used in the Review of Modelling Choices

Anderfuhren-Biget, Simon, Frédéric Varone, and David Giauque. 2014. "Policy environment and public service motivation". *Public Administration* 92 (4): 807–825.

Andersen, Lotte Bøgh, and Lene Holm Pedersen. 2012. "Public service motivation and professionalism". *International Journal of Public Administration* 35 (1): 46–57.

Andersen, Lotte Bøgh, Thomas Pallesen, and Heidi Houlberg Salomonsen. 2013. "Doing good for others and/or for society? The relationships between public service motivation, user orientation and university grading". *Scandinavian Journal of Public Administration* 17 (3): 23–44.

Andersen, Lotte Bøgh, and Lene Holm Pedersen. 2013. "Does ownership matter for employee motivation when occupation is controlled for?" *International Journal of Public Administration* 36 (12): 840–856.

Asseburg, Julia, Fabian Homberg, and Rick Vogel. 2018. "Recruitment messaging, environmental fit and public service motivation". *International Journal of Public Sector Management.*

Ballart, Xavier, and Guillem Rico. 2018. "Public or nonprofit? Career preferences and dimensions of public service motivation". *Public administration* 96 (2): 404–420.

Battaglio Jr, R Paul, and P Edward French. 2016. "Public service motivation, public management reform, and organizational socialization: Testing the effects of employment at-will and agency on PSM among municipal employees". *Public Personnel Management* 45 (2): 123–147.

Battaglio Jr, R Paul, and Salih Gelgec. 2017. "Exploring the structure and meaning of public service motivation in the Turkish public sector: a test of the mediating effects of job characteristics". *Public Management Review* 19 (8): 1066–1084.

Bellé, Nicola. 2014. "Leading to make a difference: A field experiment on the performance effects of transformational leadership, perceived social impact, and public service motivation". *Journal of Public Administration Research and Theory* 24 (1): 109–136.

Bellé, Nicola, and Paola Cantarelli. 2015. "Monetary incentives, motivation, and job effort in the public sector: An experimental study with Italian government executives". *Review of Public Personnel Administration* 35 (2): 99–123.

Brænder, Morten, and Lotte Bøgh Andersen. 2013. "Does deployment to war affect public service motivation? A panel study of soldiers before and after their service in Afghanistan". *Public Administration Review* 73 (3): 466–477.

Breaugh, Jessica, Adrian Ritz, and Kerstin Alfes. 2018. "Work motivation and public service motivation: Disentangling varieties of motivation and job satisfaction". *Public Management Review* 20 (10): 1423–1443.

Bright, Leonard. 2016. “Is public service motivation a better explanation of nonprofit career preferences than government career preferences?” *Public Personnel Management* 45 (4): 405–424.

Caillier, James Gerard. 2016a. “Do Teleworkers Possess Higher Levels of Public Service Motivation?” *Public Organization Review* 16 (4): 461–476.

— . 2016b. “Does public service motivation mediate the relationship between goal clarity and both organizational commitment and extra-role behaviours?” *Public Management Review* 18 (2): 300–318.

— . 2017. “Public service motivation and decisions to report wrongdoing in US Federal Agencies: Is this relationship mediated by the seriousness of the wrongdoing”. *The American Review of Public Administration* 47 (7): 810–825.

— . 2014. “Toward a better understanding of the relationship between transformational leadership, public service motivation, mission valence, and employee performance: A preliminary study”. *Public Personnel Management* 43 (2): 218–239.

— . 2015. “Transformational leadership and whistle-blowing attitudes: Is this relationship mediated by organizational commitment and public service motivation?” *The American Review of Public Administration* 45 (4): 458–475.

Charbonneau, Étienne, and Gregg G Van Ryzin. 2017. “Exploring the deep antecedent of public service motivation”. *International Journal of Public Administration* 40 (5): 401–407.

Chen, Chung-An, and Chih-Wei Hsieh. 2015a. “Does pursuing external incentives compromise public service motivation? Comparing the effects of job security and high pay”. *Public Management Review* 17 (8): 1190–1213.

— . 2015b. “Knowledge sharing motivation in the public sector: the role of public service motivation”. *International Review of Administrative Sciences* 81 (4): 812–832.

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.

Cheng, Kuo-Tai. 2015. “Public service motivation and job performance in public utilities”. *The International Journal of Public Sector Management* 28 (4/5): 352.

Choi, Yujin. 2017. “Work values, job characteristics, and career choice decisions: Evidence from longitudinal data”. *The American Review of Public Administration* 47 (7): 779–796.

Christensen, Robert K, Steven W Whiting, Tobin Im, Eunju Rho, Justin M Stritch, and Jungho Park. 2013. “Public service motivation, task, and non-task behavior: A performance appraisal experiment with Korean MPA and MBA students”. *International Public Management Journal* 16 (1): 28–52.

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.

Esteve, Marc, Diemo Urbig, Arjen Van Witteloostuijn, and George Boyne. 2016. “Prosocial behavior and public service motivation”. *Public Administration Review* 76 (1): 177–187.

Esteve, Marc, Arjen Van Witteloostuijn, and George Boyne. 2015. “The effects of public service motivation on collaborative behavior: Evidence from three experimental games”. *International Public Management Journal* 18 (2): 171–189.

Gamassou, Claire Edey. 2015. “What drives personnel out of public organizations?” *Public Organization Review* 15 (3): 383–398.

Gould-Williams, Julian S, Ahmed Mohammed Sayed Mostafa, and Paul Bottomley. 2015. “Public service motivation and employee outcomes in the Egyptian public sector: Testing the mediating effect of person-organization fit”. *Journal of Public Administration Research and Theory* 25 (2): 597–622.

Hamidullah, Madinah F, Gregg G Van Ryzin, and Huafang Li. 2016. “The agreeable bureaucrat: personality and PSM”. *International Journal of Public Sector Management.*

Hansen, Jesper Rosenberg, and Anne Mette Kjeldsen. 2018. “Comparing affective commitment in the public and private sectors: a comprehensive test of multiple mediation effects”. *International Public Management Journal* 21 (4): 558–588.

Holt, Stephen B. 2019. “The influence of high schools on developing public service motivation”. *International Public Management Journal* 22 (1): 127–175.

Jacobsen, Christian Bøtcher, Johan Hvitved, and Lotte Bøgh Andersen. 2014. “Command and motivation: How the perception of external interventions relates to intrinsic motivation and public service motivation”. *Public Administration* 92 (4): 790–806.

Jensen, Ulrich Thy, and Lotte Bøgh Andersen. 2015. “Public service motivation, user orientation, and prescription behaviour: Doing good for society or for the individual user?” *Public Administration* 93 (3): 753–768.

Jensen, Ulrich Thy, Lotte Bøgh Andersen, and Christian Bøtcher Jacobsen. 2019. “Only when we agree! How value congruence moderates the impact of goal-oriented leadership on public service motivation”. *Public Administration Review* 79 (1): 12–24.

Kim, Sangmook. 2017a. “Comparison of a multidimensional to a unidimensional measure of public service motivation: Predicting work attitudes”. *International Journal of Public Administration* 40 (6): 504–515.

— . 2017b. “Developing an item pool and testing measurement invariance for measuring public service motivation in Korea”. *International Review of Public Administration* 22 (3): 231–244.

— . 2012. “Does person-organization fit matter in the public-sector? Testing the mediating effect of person-organization fit in the relationship between public service motivation and work attitudes”. *Public Administration Review* 72 (6): 830–840.

— . 2017c. “National culture and public service motivation: investigating the relationship using Hofstede’s five cultural dimensions”. *International Review of Administrative Sciences* 83 (1_suppl): 23–40.

— . 2018. “Public service motivation, organizational social capital, and knowledge sharing in the Korean public sector”. *Public Performance & Management Review* 41 (1): 130–151.

— . 2010. “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.

Kim, Seung Hyun, and Sangmook Kim. 2017. “Ethnic differences in social desirability bias: Effects on the analysis of public service motivation”. *Review of Public Personnel Administration* 37 (4): 472–491.

— . 2016. “National culture and social desirability bias in measuring public service motivation”. *Administration & Society* 48 (4): 444–476.

Kjeldsen, Anne Mette. 2014. “Dynamics of public service motivation: Attraction–selection and socialization in the production and regulation of social services”. *Public Administration Review* 74 (1): 101–112.

— . 2012. “Vocational study and public service motivation: Disentangling the socializing effects of higher education”. *International Public Management Journal* 15 (4): 500–524.

Kjeldsen, Anne Mette, and Jesper Rosenberg Hansen. 2018. “Sector differences in the public service motivation–job satisfaction relationship: exploring the role of organizational characteristics”. *Review of Public Personnel Administration* 38 (1): 24–48.

Kjeldsen, Anne Mette, and Christian Bøtcher Jacobsen. 2013. “Public service motivation and employment sector: Attraction or socialization?” *Journal of Public Administration Research and Theory* 23 (4): 899–926.

Kroll, Alexander. 2014. “Why performance information use varies among public managers: Testing manager-related explanations”. *International Public Management Journal* 17 (2): 174–201.

Kroll, Alexander, and Dominik Vogel. 2014. “The PSM–leadership fit: A model of performance information use”. *Public Administration* 92 (4): 974–991.

Lee, Young-Joo, and Jin-Woo Jeong. 2015. “The link between public service motivation and volunteering: The case of South Korean civil servants”. *International Journal of Public Administration* 38 (5): 355–363.

Liu, Bangcheng, Wei Hu, and Yen-Chuan Cheng. 2015. “From the west to the east: Validating servant leadership in the Chinese public sector”. *Public Personnel Management* 44 (1): 25–45.

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.

Liu, Bangcheng, James L Perry, Xinyu Tan, and Xiaohua Zhou. 2018. “A cross-level holistic model of public service motivation”. *International Public Management Journal* 21 (5): 703–728.

Liu, Bangcheng, Thomas Li-Ping Tang, and Kaifeng Yang. 2015. “When does public service motivation fuel the job satisfaction fire? The joint moderation of person–organization fit and needs–supplies fit”. *Public Management Review* 17 (6): 876–900.

Liu, Bangcheng, Kaifeng Yang, and Wei Yu. 2015. "Work-related stressors and health-related outcomes in public service: Examining the role of public service motivation". *The American Review of Public Administration* 45 (6): 653–673.

Liu, Bangcheng, Xiaoyi Zhang, Lanying Du, and Qi Hu. 2015. "Validating the construct of public service motivation in for-profit organizations: A preliminary study". *Public Management Review* 17 (2): 262–287.

Loon, Nina Mari van, Wouter Vandenabeele, and Peter Leisink. 2015. "On the bright and dark side of public service motivation: The relationship between PSM and employee wellbeing". *Public Money & Management* 35 (5): 349–356.

Luu, Tuan. 2018. "Discretionary HR practices and proactive work behaviour: the mediation role of affective commitment and the moderation roles of PSM and abusive supervision". *Public Management Review* 20 (6): 789–823.

Meyer, Renate E, Isabell Egger-Peitler, Markus A Höllerer, and Gerhard Hammerschmid. 2014. "Of bureaucrats and passionate public managers: Institutional logics, executive identities, and public service motivation". *Public Administration* 92 (4): 861–885.

Ngaruiya, Katherine M, Anne-Lise Knox Velez, Richard M Clerkin, and Jami Kathleen Taylor. 2014. "Public service motivation and institutional-occupational motivations among undergraduate students and ROTC cadets". *Public Personnel Management* 43 (4): 442–458.

Nowell, Branda, Anne M Izod, Katherine M Ngaruiya, and Neil M Boyd. 2016. "Public service motivation and sense of community responsibility: Comparing two motivational constructs in understanding leadership within community collaboratives". *Journal of Public Administration Research and Theory* 26 (4): 663–676.

Olsen, Asmus Leth, Frederik Hjorth, Nikolaj Harmon, and Sebastian Barfort. 2019. "Behavioral dishonesty in the public sector". *Journal of Public Administration Research and Theory* 29 (4): 572–590.

Park, Seejeen. 2014. "Motivation of public managers as raters in performance appraisal: Developing a model of rater motivation". *Public Personnel Management* 43 (4): 387–414.

Pedersen, Lene Holm. 2014. "Committed to the public interest? Motivation and behavioural outcomes among local councillors". *Public Administration* 92 (4): 886–901.

Pedersen, Mogens Jin, Justin M Stritch, and Gabel Taggart. 2017. "Citizen perceptions of procedural fairness and the moderating roles of 'belief in a just world'and 'public service motivation'in public hiring". *Public Administration* 95 (4): 874–894.

Piatak, Jaclyn Schede. 2016. "Public service motivation, prosocial behaviours, and career ambitions". *International Journal of Manpower* 37 (5): 804–821.

Potipiroon, Wisanupong, and Michael T Ford. 2017. "Does public service motivation always lead to organizational commitment? Examining the moderating roles of intrinsic motivation and ethical leadership". *Public Personnel Management* 46 (3): 211–238.

Quratulain, Samina, and Abdul Karim Khan. 2015. "How does employees' public service motivation get affected? A conditional process analysis of the effects of person–job fit and work pressure". *Public Personnel Management* 44 (2): 266–289.

Rayner, Julie, Vaughan Reimers, and Chih-Wei Chao. 2018. "Testing an International Measure of Public Service Motivation: Is There Really a Bright or Dark Side?" *Australian Journal of Public Administration* 77 (1): 87–101.

Ritz, Adrian. 2015. "Public service motivation and politics: Behavioural consequences among local councillors in Switzerland". *Public Administration* 93 (4): 1121–1137.

Sanabria-Pulido, Pablo. 2018. "Public service motivation and job sector choice: Evidence from a developing country". *International Journal of Public Administration* 41 (13): 1107–1118.

Schott, Carina, Daphne D Van Kleef, and Trui PS Steen. 2018. "The combined impact of professional role identity and public service motivation on decision-making in dilemma situations". *International Review of Administrative Sciences* 84 (1): 21–41.

Schwarz, Gary, Alexander Newman, Brian Cooper, and Nathan Eva. 2016. "Servant leadership and follower job performance: The mediating effect of public service motivation". *Public Administration* 94 (4): 1025–1041.

Shim, Dong Chul, and Hyun Hee Park. 2019. "Public service motivation in a work group: Role of ethical climate and servant leadership". *Public Personnel Management* 48 (2): 203–225.

Shim, Dong Chul, Hyun Hee Park, and Tae Ho Eom. 2017. "Street-level bureaucrats' turnover intention: does public service motivation matter?" *International Review of Administrative Sciences* 83 (3): 563–582.

Shrestha, Arjun Kumar, and Ajaya Kumar Mishra. 2015. "Interactive effects of public service motivation and organizational politics on Nepali civil service employees' organizational commitment". *Business Perspectives and Research* 3 (1): 21–35.

Song, Miyeon, Illoong Kwon, Seyeong Cha, and Naon Min. 2017. "The effect of public service motivation and job level on bureaucrats' preferences for direct policy instruments". *Journal of Public Administration Research and Theory* 27 (1): 36–51.

Tepe, Markus. 2016. "In public servants we trust?: a behavioural experiment on public service motivation and trust among students of public administration, business sciences and law". *Public Management Review* 18 (4): 508–538.

Van Loon, Nina Mari. 2017. "Does context matter for the type of performance-related behavior of public service motivated employees?" *Review of public personnel administration* 37 (4): 405–429.

— . 2016. "Is public service motivation related to overall and dimensional work-unit performance as indicated by supervisors?" *International Public Management Journal* 19 (1): 78–110.

Van Loon, Nina Mari, Wouter Vandenabeele, and Peter Leisink. 2017. "Clarifying the relationship between public service motivation and in-role and extra-role behaviors: The relative contributions of person-job and person-organization fit". *The American Review of Public Administration* 47 (6): 699–713.

Van Witteloostuijn, Arjen, Marc Esteve, and George Boyne. 2017. "Public sector motivation ad fonts: Personality traits as antecedents of the motivation to serve the public interest". *Journal of public administration research and theory* 27 (1): 20–35.

Ward, Kevin D. 2014. "Cultivating public service motivation through AmeriCorps service: A longitudinal study". *Public Administration Review* 74 (1): 114–125.

Wright, Bradley E, Robert K Christensen, and Kimberley Roussin Isett. 2013. "Motivated to adapt? The role of public service motivation as employees face organizational change". *Public Administration Review* 73 (5): 738–747.

Wright, Bradley E, Robert K Christensen, and Sanjay K Pandey. 2013. "Measuring public service motivation: Exploring the equivalence of existing global measures". *International Public Management Journal* 16 (2): 197–223.

Wright, Bradley E, Shahidul Hassan, and Robert K Christensen. 2017. "Job choice and performance: Revisiting core assumptions about public service motivation". *International Public Management Journal* 20 (1): 108–131.

Wright, Bradley E, Shahidul Hassan, and Jongsoo Park. 2016. "Does a public service ethic encourage ethical behaviour? Public service motivation, ethical leadership and the willingness to report ethical problems". *Public Administration* 94 (3): 647–663.

Yeo, Jungwon. 2016. "Recent Administrative and Managerial Practices and Public Service Motivation: Evidence from Seoul City Government, South Korea". *International Journal of Public Administration* 39 (3): 216–225.