CHAPTER 24

Interpreting Survey Findings

Can Survey Results Be Compared across Organizations and Countries?

Robert Lipinski, Jan-Hinrik Meyer-Sahling, Kim Sass Mikkelsen, and Christian Schuster

SUMMARY

With the rise in worldwide efforts to understand public administration by surveying civil servants, issues of survey question comparability become paramount. Surveys can rarely be understood in a void but rather require benchmarks and points of reference. However, it is not clear whether survey questions, even when phrased and structured in the same manner, measure the same concepts in the same way and, therefore, can be compared. For multiple reasons, including work environment, adaptive expectations, and cultural factors, different people might understand the same question in different ways and adjust their answers accordingly. This might make survey results incomparable, not only across countries but also across different groups of civil servants within a national public administration. This chapter uses results from seven public service surveys from across Europe, Latin America, and South Asia to investigate the extent to which the same survey questions measure the same concepts similarly—that is, are measurement invariant—using as an example questions related to *transformational leadership*. To ascertain measurement invariance, models of a hypothesized relationship between questions measuring transformational leadership are compared across countries, as well as along gender, educational, and organizational lines within countries. Solid evidence of metric invariance and tentative evidence of scalar invariance is found in cross-country comparisons. Moreover, factor loadings can be judged equal (*metric invariance*) across gender, education level, and organization in most countries, as can latent factor means (*scalar invariance*). Our results suggest that groups of public servants within countries—delineated, for instance, by gender, education, or organization—can typically be benchmarked without invariance concerns. Across countries, evidence for valid benchmarking—that is, scalar invariance—is strongest for countries in similar regions and at similar income levels. It is weaker—though still suggestive—when comparing all countries in the sample. Our chapter concludes that less culturally contingent concepts may be plausibly benchmarked with care across countries.

Robert Lipinski is a consultant in the World Bank's Development Impact Evaluation (DIME) Department. Jan-Hinrik Meyer-Sahling is a professor at the University of Nottingham. Kim Sass Mikkelsen is an associate professor at Roskilde University. Christian Schuster is a professor at University College London.

ANALYTICS IN PRACTICE

- Many theoretical insights and practical lessons from surveys of civil servants depend on the ability to draw comparisons between countries and demographic groups. This chapter focuses on the comparability of the concept of *transformational leadership* across different contexts and groups—a premise known as *measurement invariance*. Transformational leadership measures the extent to which managers lead by setting a good example, making employees proud, and generating enthusiasm about an organization's mission.
- Equality in the understanding of a single overarching concept, such as transformational leadership, across different countries and groups can be conceptualized in three ways. The same concept can be measured by the same set of questions (*configural invariance*), those questions can have the same strength of a relationship with the underlying concept (*metric invariance*), and the concept can have the latent mean structure (*scalar invariance*).
- When comparing survey measures and concepts across countries, practitioners should consider the extent to which different cultural interpretations of a concept (such as leadership), different social-desirability biases, different pressures in the work environment, and even differences in language may lead to differences in survey means across countries that do not reflect substantive differences in the underlying concept (such as the quality of leadership).
- When empirically assessing the measurement invariance (and thus the cross-country comparability) of a concept that is arguably culturally specific—transformational leadership—we find that cross-country comparisons can be undertaken, although with caution. There is evidence that the concept of transformational leadership is understood in a comparable way across the seven countries included in the analyses. As we find suggestive evidence that cross-country comparisons are possible with even a relatively culturally contingent concept (leadership), cross-country comparisons of more factual questions (for example, "Did you have a performance evaluation last year?") are plausibly often possible in a valid manner.
- Grouping countries by region and income level removes many of the differences across countries. This suggests that comparisons between countries at similar income levels and in the same world regions can be made with greater confidence.
- Within-country comparisons of transformational leadership suffer from fewer concerns about lack of comparability. Empirically, we find that they can be reliably made across public servants of different genders and education levels, and in different institutions.

INTRODUCTION

Surveys of civil servants provide insights into core parts of the public administration production function—such as the quality of management and the attitudes (for example, motivation) of employees. As argued by Rogger and Schuster in chapters 1–3 of *The Government Analytics Handbook*, these determinants of public sector productivity are difficult to measure accurately with other data sources. Survey results are typically presented as percentages of public servants who evaluate favorably dimensions of their work environment, management, or themselves—for instance, the percentage of public servants who recommend their organization as a great place to work, or the percentage of public servants who evaluate the leadership of their superior favorably.

How can governments know whether certain percentages—such as 75 percent of public servants who are satisfied with their jobs—are strengths or weaknesses of their public service? Interpreting survey results—and understanding areas for development in the public service—is often greatly aided by comparison. By benchmarking themselves with other countries on the same survey response, governments can understand where their strengths and weaknesses lie. This is one of the founding motivations of the Global Survey of Public Servants (GSPS) initiative (Fukuyama et al. 2022). Similarly, benchmarking internally between groups of public servants—for example, by gender, education, or institution—can help governments understand where, inside government, strengths and weaknesses lie.

However, such benchmarking presupposes comparability in measurement and the survey response process. In other words, it presupposes that respondents understand concepts—such as leadership, motivation, and satisfaction—in the same manner across different countries, government institutions, or groups (for example, men and women) in public service, and that they face similar biases (for example, social-desirability bias) when responding to survey questions. If the same concepts mean different things to different public servants or trigger different response biases in different public servants, valid comparisons are no longer possible, as differences in means might stem from differences in understanding or bias rather than differences in the underlying concept (for example, differences in actual work motivation).

Many public service survey questions are filtered through cultural factors (for example, "My direct superior leads by setting a good example"), individual-level characteristics, like gender ("I am paid at least as well as colleagues who have job responsibilities similar to me"), or both ("I feel sympathetic to the plight of the underprivileged"). If that is the case, then the survey measure lacks *measurement invariance*, which is "a property of a measurement instrument (in the case of survey research, a questionnaire), implying that the instrument measures the same concept in the same way across various subgroups of respondents" (Davidov et al. 2014, 58).

Past research suggests that measurement invariance might affect some measures in surveys of public servants and, in particular, public service motivation (PSM) (Kim et al. 2013; Mikkelsen, Schuster, and Meyer-Sahling 2020). However, what it *means* to be motivated to serve the public in all its dimensions—such as commitment to public values or compassion—is, arguably, highly dependent on cultural factors. As such, concerns about PSM's lack of cross-country comparability might not travel to other survey questions with less-cultural and more-factual content.

To assess this empirically, this chapter assesses what is arguably a key determinant of public administration effectiveness: the quality of leadership and, in particular, the concept of *transformational leadership*, a style of leadership that inspires and motivates subordinates to go beyond their self-interest and expectation of pecuniary rewards to achieve their goals and an organization's targets (Jensen et al. 2019; Pearce et al. 2002). Transformational leadership has been found to positively affect performance in public sector organizations across multiple contexts (Hameduddin and Engbers 2021; Pandey et al. 2016; Schuster et al. 2020).

Methodologically, we follow Mikkelsen, Schuster, and Meyer-Sahling (2020, 740) and undertake a measurement-invariance analysis given that "systematic cross-cultural and cross-national measurement-invariance analyses are central to gauge the comparability and generalizability." We apply the measurement-invariance analysis to an original seven country survey of public servants, in which transformational leadership is measured with exactly the same measurement scale across countries. We assess measurement invariance across countries and within countries across government institutions, as well as across public servants with different genders and education levels.

Our chapter is organized as follows. The chapter begins with a review of the measurement-invariance literature, with a particular focus on its application in the field of public service surveying and on the concept of transformational leadership within the civil service. It then proceeds to describe the approach taken to analyze the measurement invariance of the concept of transformational leadership, including the data set used and the method of analysis: multigroup confirmatory factor analysis (MGCFA). After that, we present our results—first, for cross-country comparisons and then for within-country comparisons, for civil servants grouped by gender, education level, and organization. We then discuss the theoretical and practical implications of our results and conclude.

LITERATURE REVIEW

The Concept of Measurement Invariance

It is common for surveys to aggregate individual questions into larger, overarching constructs. For example, the United States Office of Personnel Management (OPM) Federal Employee Viewpoint Survey (FEVS) calculates three subindexes pertaining to some key aspects of public service functioning ("leaders lead," "supervisors," and "intrinsic work experience"), each composed by averaging positive responses to five survey questions. These are, in turn, aggregated into an "employee engagement index" (OPM 2019). To take another example, the United Kingdom's Civil Service People Survey also calculates an "employee engagement index," tabulated over five questions selected based on factor analysis from pilot surveys (Cabinet Office 2019). Despite similarities in their names and their high-income, English-speaking country settings, however, these two measures cannot be directly compared with each other, due to differences in wording and survey methodology. However, another long-standing concern of survey researchers is the possibility that even exactly the same questions can be interpreted differently by various groups of respondents. Engagement measured with the same battery of questions could still be conceived differently by civil servants in the United States and the United Kingdom due to cultural differences, institutional context, or socioeconomic factors.

Therefore, in order to meaningfully compare a statistical construct, like engagement, motivation, or leadership, and related statistical quantities, like means and regression coefficients, across different groups (or time periods), the construct should first be tested for measurement invariance. Demonstrating the measurement invariance (sometimes also termed equivalence) of a given construct entails showing that it is interpreted in a comparable manner by different sets of respondents. In contrast, "measurement *non*-invariance suggests that a construct has a different structure or meaning to different groups or on different measurement occasions in the same group, and so the construct cannot be meaningfully tested or construed across groups or across time" (Putnick and Bornstein 2016, 71; emphasis added).

Three basic levels of measurement invariance are usually distinguished: *configural*, *metric*, and *scalar* (Vandenberg and Lance 2000). They represent progressively stricter tests for comparability between groups. Figure 24.1, below, provides a schematic representation of the generalized idea behind these concepts by illustrating how each of them hypothesizes the relationship between manifest variables and underlying latent constructs. A more detailed visualization is provided by figures L.1, L.2, and L.3 in appendix L. They demonstrate different levels of invariance using examples of models of transformational leadership in public service that are analyzed throughout this chapter.

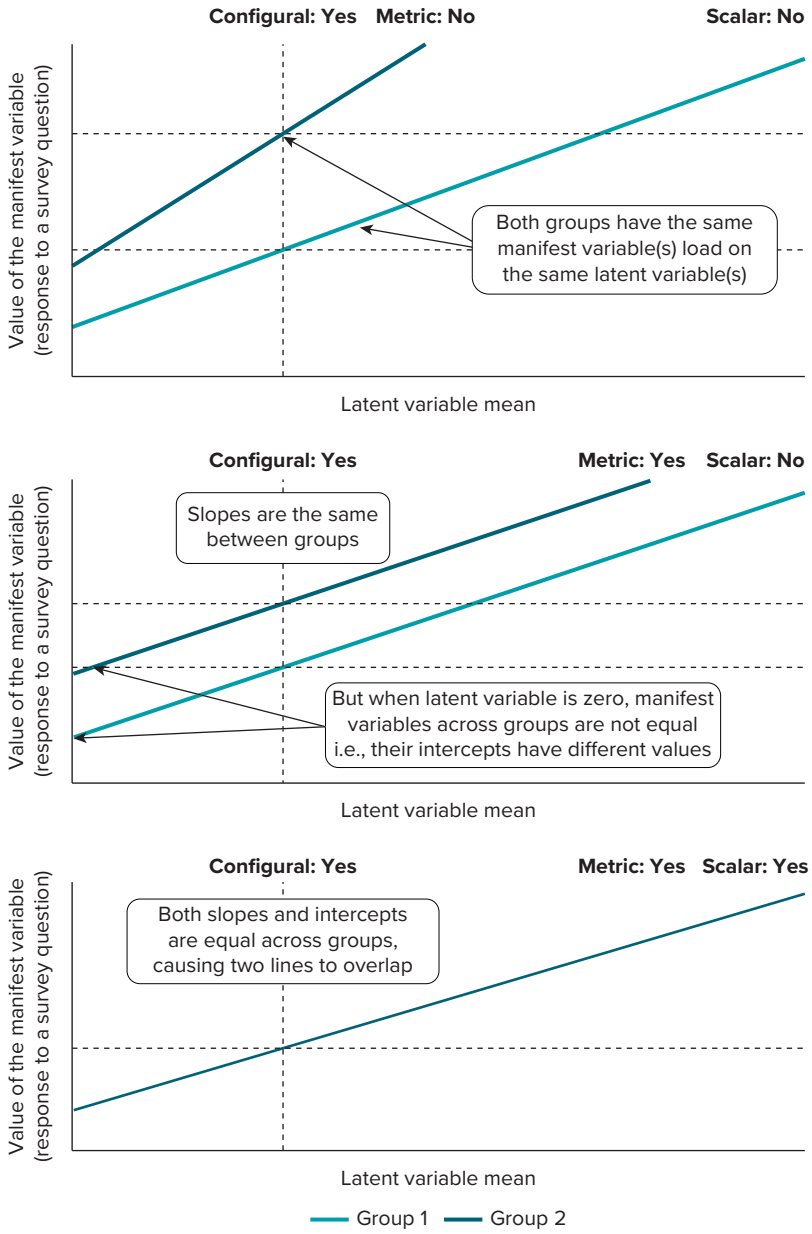
Configural Invariance

In the first step toward establishing the comparability of a statistical concept, it needs to be ascertained that it has the same *factor structure* across all groups being compared. This means that in all groups, the same sets of questions are linked to the same sets of underlying constructs, typically termed *latent factors* or *variables* (Kim et al. 2013). This is schematically represented by the top panel of figure 24.1 and, on the example of transformational leadership, by figure L.1 in appendix L. If, in both groups of interest, it can be shown that a model with all observed survey questions loading onto a single latent variable fits the data well, then configural invariance is deemed to hold (upper panel of figure L.1). However, if this hypothesized model is found not to fit the data, then configural invariance cannot be said to hold. One example of such a situation could be where data from one group display a two-latent-factor structure, as in the bottom panel of figure L.1.

Metric Invariance

Once the same structure of the items and factors is confirmed across groups, researchers might turn their attention to the equality of *factor loadings* between the groups. Factor loadings can be understood

FIGURE 24.1 Schematic Visual Representation of the Three Levels of Measurement Invariance: Configural, Metric, and Scalar



Source: Adapted from Cieciuch et al. 2019, 179.
Note: "The X axis represents the latent variable mean; the Y axis represents the response to a survey question item measuring the latent variable. The diagonal represents the function relation between the latent variable and the response to the survey question item in two countries (in unstandardized terms)" (Cieciuch et al. 2019, 179). Here, *countries* is replaced by the more generic term *groups*.

as measures of the strength of the relationship between the observed survey items and the latent factors. Confirming metric invariance is a pre-requisite that "ensures that *structural regression estimates* are comparable across groups" (Mikkelsen, Schuster, and Meyer-Sahling 2020, 4; emphasis added). This is because, in metric-invariant models, differences between survey items are linked to differences in the underlying latent-factor models in the same fashion across all the groups included in the analyses (Steenkamp and Baumgartner 1998). The equal slopes of the lines in the middle panel of figure 24.1 demonstrate this point. In other words, one unit change in the *x*-axis value of the latent variable is associated

with a change in manifest variable values that is the same for both groups. Consequently, only with metric invariance can regression with observed survey items be compared in a meaningful manner (Hong, Malik, and Lee 2003). This focus of metric invariance on the equality of factor loadings across groups is also shown in figure L.2 in appendix L.

Scalar Invariance

To ensure scalar invariance, not only factor loadings but also the means of the *item intercepts* must be shown to be equal between groups (Vandenberg and Lance 2000). Even when factor loadings suggest that the latent constructs have the same impact upon the value of observed items in all the groups considered, as in the definition of metric invariance, it is still possible for groups to have different values of the intercept—that is, the value of observed items when the latent variable is zero—due to some unobservable characteristics.

Only when the intercepts are the same in all groups, as in the bottom panel of figure 24.1, can the model be said to be scalar invariant. This is also the situation presented in figure L.3 in appendix L. Thus, establishing scalar invariance should precede any attempt at comparing the latent means and intercepts of observed items between the groups. Only once it is established can it can be assumed that "cross-national differences in the means of the observed items are due to differences in the means of the underlying construct(s)" (Steenkamp and Baumgartner 1998, 80).

Measurement Invariance

Although first developed in the mid-20th century in the field of psychology (see, for example, Meredith 1964; Struening and Cohen 1963), measurement invariance has since become a concern in multiple other disciplines. In the field of education, researchers have applied it to better understand the comparability of concepts such as the time management of US undergraduate students (Martinez 2021) or the different subscales of school climate measured in the Georgia School Climate Survey (La Salle, McCoach, and Meyers 2021). The Organisation for Economic Co-operation and Development (OECD) has used it to gauge the comparability of latent factors measured by the Programme for International Student Assessment (PISA) and several other cross-country surveys (Van De Vijveri et al. 2019). It has also been analyzed in other contexts as diverse as consumer research (De Jong, Steenkamp, and Fox 2007) and sociology—for example, to better understand concepts such as attitudes toward granting citizenship rights in the International Social Survey Program (Davidov et al. 2018) and German adolescents' life attitudes (Seddig and Leitgöb 2018). Given its importance and the widespread academic interest in it, public administration researchers, in the face of increasing surveying efforts, have also come to analyze measurement invariance in public service surveys.

Measurement Invariance in Public Service Surveys

In recent years, multiple researchers have emphasized the importance of studying public administration from a comparative perspective, both to improve researchers' theoretical understanding and to draw practical lessons (Fitzpatrick et al. 2011; Jreisat 2005). At the same time, surveys have become one of the key methods used to better understand public administration. As emphasized throughout this part of the *Handbook*, surveys allow researchers and policy practitioners to gain insights into dimensions of public administration's functioning that would otherwise be unmeasurable. Concepts such as job satisfaction or attitudes toward management could scarcely be gauged otherwise. Surveys can also be used to anonymously ask about aspects of civil servants' work that might otherwise not be talked about—such as perceptions of corruption or workplace harassment. However, the comparability of survey results—both across countries and across demographic groups within the civil service—cannot be taken for granted. Challenges to comparability stem from several sources, like differences in the mode of survey delivery (see chapter 19) or perceived question sensitivity (see chapter 22). Another obstacle is the different phrasing of questions—an issue that initiatives such as the GSPS have recently begun to address (Fukuyama et al. 2022). However, even with all these problems solved, it is not certain that the same survey concepts would be understood in the same way by different groups of civil servants.

This explains the recent turn of several public administration scholars toward analyzing measurement invariance across public service surveys. Kim et al. (2013) test for measurement invariance in PSM—one of the frequently recurring parts of many public service survey questionnaires, which aims to measure respondents' motivation and willingness to serve society. The authors use a PSM index containing questions asked using a 1–5 Likert scale. The tests for configural invariance suggest that PSM has the same structure in 8 out of the 12 countries studied. However, neither metric nor scalar invariance can be detected, meaning that the construct has a different meaning and different levels across countries (apart from the sample restricted to the culturally similar Australia, the United Kingdom, and the United States, for which metric invariance can be detected). Mikkelsen, Schuster, and Meyer-Sahling (2020) expand these results by using a more diverse sample of countries and larger sample sizes within countries. Using survey results from over 23,000 civil servants across 10 countries, they demonstrate that a 16-item PSM scale displays first- and second-order partial metric invariance, apart from the case of two Asian countries studied (Bangladesh and Nepal). Still, their study finds that PSM levels cannot be compared across countries due to a lack of scalar invariance.

The Concept of Transformational Leadership

In the face of the expansion of measurement-invariance studies within the field of public administration, a relative lack of attention to concepts other than PSM can be discerned. Although PSM is of clear importance and is commonly measured, many other dimensions of work affect civil servants' performance and are regularly included in public service surveys. One key concept, measured in some form in virtually every public service survey, is leadership. Its measurement is most frequently based on past research, scales, and wording from the management science and psychology literature (Tummers and Knies 2016). In particular, the idea of *transformational leadership* has gained traction with public administration researchers (see, for example, Kroll and Vogel 2014; Pandey et al. 2016). It was first developed in the 1970s by Downton (1973) and more fully by Burns (1978), who applied it, together with the contrasting idea of *transactional leadership*, to study political leaders. Whereas transactional leadership is conceived as a leadership style focused on tangible benefits obtained via exchange between a leader and followers (for example, jobs for votes), transformational leadership is chiefly focused on motivating and engaging potential followers to move in a desired direction by conveying a sense of mission, employing compelling argumentation, and using one's own example. Under transformational leadership, followers are inspired to maximize their performance and achieve set goals for the sake of "higher level needs such as self-actualization" (Pearce et al. 2002, 281), attention, and personal development (Nguyen et al. 2017). Bass (1985) extended both conceptions of leadership to the management of organizations. Since then, transformational leadership has been found to be one of the key factors explaining improvements in many dimensions of performance in multiple settings in the private sector, including increased agreement on strategic goals in a large Israeli telecommunications firm (Berson and Avolio 2004), satisfaction with supervisors in Turkish boutique hotels (Erkutlu 2008), and knowledge management in Spanish firms (García-Morales, Lloréns-Monte, and Verdú-Jover 2008). A meta-analysis of 113 primary studies on the topic by Wang et al. (2011) finds transformational leadership to be associated with better performance across the individual, team, and organizational levels.

Moreover, a recent meta-analysis of the PSM and leadership literature, conducted by Hameduddin and Engbers (2021), has found that 50 percent ($n = 20$) of publications concerned with leadership rely on the concept of transformational leadership, making it the most common conceptualization of leadership by public administration scholars. Following this approach, Park and Rainey (2008) establish a positive relationship between transformational leadership and outcomes such as job satisfaction, quality of work, and perceived performance across US federal agencies. Pandey et al. (2016) find its direct and indirect impacts on normative public values. Donkor, Sekyere, and Oduro (2022) further find that higher transformational leadership is linked to higher organizational commitment across 16 Ghanaian public sector organizations. In a survey of over 21,000 civil servants in Chile, Schuster et al. (2020) similarly find transformational leadership to be correlated with higher job satisfaction, motivation, and engagement. Hameduddin and Engbers' (2021) review

of 40 studies finds a link between transformational leadership and PSM—a relationship that holds across a diverse set of countries analyzed.

Transformational leadership was therefore chosen as the survey instrument of focus in the present chapter because of its solid theoretical development, extensive academic research pedigree, and practical importance for public sector performance. Two further reasons can be adduced to explain this choice. First, it is a concept that can usually be mapped onto a single underlying construct. In other words, survey questions about transformational leadership are all aimed at measuring different but related aspects of the same latent factor. This is often not the case with many other sections of public service surveys, like salaries or performance management, which measure many divergent subdimensions—including administrative (for example, salary amount and participation in performance evaluations), motivational (for example, satisfaction with salary and usefulness of performance evaluations), and ethical (for example, salary and performance evaluations' fairness) subdimensions.

Second, there exists a relative imbalance between the large number of studies relying on measures of transformational leadership in the public sector and the lack of research investigating the measurement invariance of this concept. To the best of the authors' knowledge, the only analysis of measurement invariance focused on transformational (and transactional) leadership is a paper by Jensen et al. (2019). However, it presents only a limited test of measurement invariance for transformational leadership, as it focused on full configural and metric invariance, without tests of partial metric invariance or scalar invariance. Jensen et al. (2019) also do not engage in cross-country or cross-cultural analysis of invariance because their sample is composed of respondents from Denmark. The authors focus on invariance across time, sector (including public vs. private), and randomized training groups but not demographic variables, like gender or education, or organizations within the public sector—a focus of the present chapter. Thus, although transformational leadership has gained a well-established position within the public administration literature, only limited attention has been paid to testing the measurement invariance of this concept, which provides the rationale for the analyses contained in the pages below.

METHODOLOGY

Data Set

The data used for the analysis in this chapter come from the GSPS initiative. The GSPS is a combined effort of researchers at the World Bank's Bureaucracy Lab, University College London (UCL), the University of Nottingham, and Stanford University that aims to better understand the attitudes and behaviors of civil servants around the globe. Part of the GSPS is focused on making public administration survey questionnaires more comparable. It strives to achieve this by developing and promoting the inclusion of a "core" survey module, which would ask the same set of questions about the principal dimensions of civil service work, such as job satisfaction, work motivation, and leadership, to all the civil servants surveyed.

Seven public service surveys are included in the analyses below. They come from the following countries: Albania, Bangladesh, Brazil, Chile, Estonia, Kosovo, and Nepal.[1] Together, the surveys gathered responses from over 21,000 civil servants. Surveys were delivered both online and in person between 2017 and 2018 and included an extensive set of questions pertaining to multiple aspects of civil service functioning.[2] Importantly for present purposes, the phrasing of questions was exactly the same across countries. In order to ensure that respondents' understanding of the questions would remain unaffected by translation into local languages, the questions were pretested using cognitive interviews with civil servants and iteratively revised (Mikkelsen, Schuster, and Meyer-Sahling 2020). Moreover, each survey strove to include a comparable sample of respondents—that is, central government civil servants who perform general administrative duties.[3] Due to incomplete personnel records on civil servants, the samples are not fully representative. Furthermore, in the in-person surveys, informal quota sampling and in-person surveys based on information from

individual public administration organizations were used (see Mikkelsen, Schuster, and Meyer-Sahling 2020). When possible, the demographics of the survey samples were compared to servicewide values (see table 24A.1), and those comparisons reveal broadly aligned values.

The final advantage of the present choice of surveys is that they represent a diverse set of regional and economic groupings: from South America through Europe to Asia, and from lower-middle-income countries, like Bangladesh and Nepal, through upper-middle-income Albania, Brazil, and Kosovo to high-income Chile and Estonia (see table 24.1). This allows analyses in this chapter to not only focus on differences between groups within each civil service but also to compare invariance across the cross-cultural contexts of different regions and countries.

In the present sample of civil servant surveys, the concept of transformational leadership was measured using the level of agreement with the following three questions, all starting with the prompt "To what extent do you agree with the following statements?":

1. My direct superior articulates and generates enthusiasm for my organization's vision and mission (abbreviated as *enthusiasm*).
2. My direct superior leads by setting a good example (abbreviated as *good example*).
3. My direct superior says things that make employees proud to be part of this organization (abbreviated as *pride*).

The responses were measured using a 1–5 Likert scale, where 1 signified "strongly disagree" and 5 "strongly agree." The basic statistics on each of the variables are presented in table 24.2. A majority of the respondents agree with the question prompts, confirming that their direct superiors generate enthusiasm about the organization's vision and mission, lead by setting a good example, and make them proud to be a part of the organization. Correlations between the three variables are also very high (>0.75), which could be interpreted as an early indication that they indeed measure one underlying concept of transformational leadership.

TABLE 24.1 Summary of the Seven Public Servant Surveys Used in the Chapter

	Albania	Bangladesh	Brazil	Chile	Estonia	Kosovo	Nepal
Respondents	3,690	1,049	3,992	5,742	3,555	2,465	1,249
Response rate	47%	Convenience sample	11%	37%	25%	14%	Convenience sample
Mode of delivery	Online	In-person	Online	Online	Online	Online	In-person
Year	2017	2017–18	2018	2016–17	2017	2017	2017–18
Language	Albanian	English, Bangla	Portuguese	Spanish	Estonian	Albanian, Serbian	English, Nepali
Report	Meyer-Sahling et al. (2018d)	Meyer-Sahling et al. (2019)	Pereira et al. (2021)	Schuster et al. (2017)	Meyer-Sahling et al. (2018a)	Meyer-Sahling et al. (2018b)	Meyer-Sahling et al. (2018c)
Region[a]	ECA	South Asia	LAC	LAC	ECA	ECA	South Asia
Income group[a]	Upper-middle income	Lower-middle income	Upper-middle income	High income	High income	Upper-middle income	Lower-middle income
GDP per capita (current US$)[a]	$5,246	$1,967	$6,797	$13,232	$23,027	$4,347	$1,155

Source: Original table for this publication.
Note: ECA = Europe and Central Asia; LAC = Latin America and the Caribbean.
a. Based on World Bank data and groupings.

TABLE 24.2 Basic Statistics on the Three Questions Aiming to Measure Transformational Leadership

Statistic	Variable		
	Enthusiasm	Good example	Pride
Mean	3.59	3.74	3.40
Median	4.00	4.00	4.00
SD	1.29	1.27	1.28
Skew	−0.63	−0.81	−0.42
Kurtosis	2.29	2.61	2.11
Corr. with *enthusiasm*	1.00	0.80	0.84
Corr. with *good example*	0.80	1.00	0.79
Corr. with *pride*	0.84	0.79	1.00

Source: Original table for this publication.
Note: All variables are measured on a 1–5 Likert scale, where higher values indicate greater agreement. The values shown in the table are aggregated across countries. SD = standard deviation.

Measuring Invariance

As discussed more broadly in the literature review section, invariance can be measured on three key levels: configural, metric, and scalar. These levels of invariance are tested here using MGCFA. This has been the main method of testing measurement invariance in the past three decades (see, for example, Hofman, Mathieu, and Jacobs 1990; Mikkelsen, Schuster, and Meyer-Sahling 2020; Putnick and Bornstein 2016). It is carried out by setting progressively stricter constraints upon the parameters of the model being evaluated. First, the same model structure is imposed on all groups tested. If the model fit proves satisfactory (see the subsection below for criteria on this), metric invariance is tested by restricting factor loadings to be equal across groups. If the results from the comparison of model fit show that the constrained model is not performing significantly worse than the unconstrained one, then metric invariance can be inferred. Upon finding evidence of metric invariance, the means of the latent construct can be set to be equal across the groups, and, if this extra restriction also does not result in significantly worse model fit, then scalar invariance can be ascertained.

The first set of MGCFA tests pertains to measurement invariance across the seven countries included in the study. First, models for a set of countries grouped by region and income level are fit, before moving to full cross-country models. The results therefore demonstrate the extent to which national context determines how civil servants understand the concept of transformational leadership. The second set of analyses turns toward demographic groups within countries and evaluates whether respondents of different genders (female vs. male), education levels (below university vs. university), and organizations within the public administration interpret the questions about transformational leadership in the same manner. These two levels of analysis—inter- and intracountry—have been the key focus of measurement-invariance research (Vandenberg and Lance 2000).

Model Fit Indexes

To compare the progressively more restricted measurement-invariance models, one has to calculate how well they fit the data. Three measures are relied upon for this purpose. The first one is chi-square (χ^2). This is a likelihood ratio test that calculates how well the specified model and the associated expected distributions fit the observed data distributions. The χ^2 value, combined with the model's degrees of freedom, can be used

to calculate the *p*-value—the likelihood that the observed deviation from the perfect model is due to chance. However, researchers are in agreement that, because the mathematical formula for its derivation is dependent on the sample size (*N*), this statistic is highly sensitive in large samples and might show statistically significant differences in model fit even when only small deviations from perfect fit are present (Byrne, Shavelson, and Muthèn 1989; Cheung and Rensvold 2002; French and Finch 2006; Putnick and Bornstein 2016).

For this reason, two further fit indexes are consulted when comparing model fit. One is the comparative fit index (CFI). Its value is scaled between 0 and 1 and is specifically designed to deal with the limitations of χ_2, including its oversensitiveness in large samples (Bentler 1990). The model might be assumed to fit well already when the CFI is above 0.90 (Cheung and Rensvold 2002), but a more restrictive threshold of 0.95 is typically used (Hooper, Coughlan, and Mullen 2008; Hu and Bentler 1999). However, in the measurement-invariance literature, if restricting model parameters leads to a decrease in the CFI of more than 0.01, the invariance is typically rejected (Cheung and Rensvold 2002).

The third and final fit index consulted throughout the analyses is the standardized root mean squared error (SRMR) (see Bentler 1995). The SRMR is calculated on a range from 0 to 1 and can be viewed "as the average standardized residual covariance" of the model variables (Shi, Maydeu-Olivares, and Rosseel 2020, 2). It can range from 0 to infinity, and, typically, absolute SRMR values below 0.05 are indicative of good model fit, although values up to 0.08 are deemed satisfactory (Hu and Bentler 1999). When the fit of models is compared for invariance, increases in the SRMR of more than 0.03 and 0.01 are taken as signaling significant model deterioration in metric- and scalar-invariance models, respectively (Chen 2007). Given the large sample sizes used here, the concern with the overrejection of invariant models by the SRMR raised by Chen (2007) is largely ameliorated.[4]

Therefore, when discussing model fit below, whether in absolute terms or when comparing its fit to another model, changes (indicated with Δ) in all fit indexes are reported (the *p*-value of $\Delta\chi^2$, ΔCFI, and ΔSRMR). The models are estimated in RStudio using the lavaan::cfa() function. Given a nonsymmetrical distribution and the ordinal nature of the data (see table 24.2), a diagonally weighted least squares (DWLS) estimator is used for model estimation (Li 2016; Rosseel 2012). Comparisons of model fit (χ_2, CFI, and SRMR values), are made using the semTools::compareFit() function.

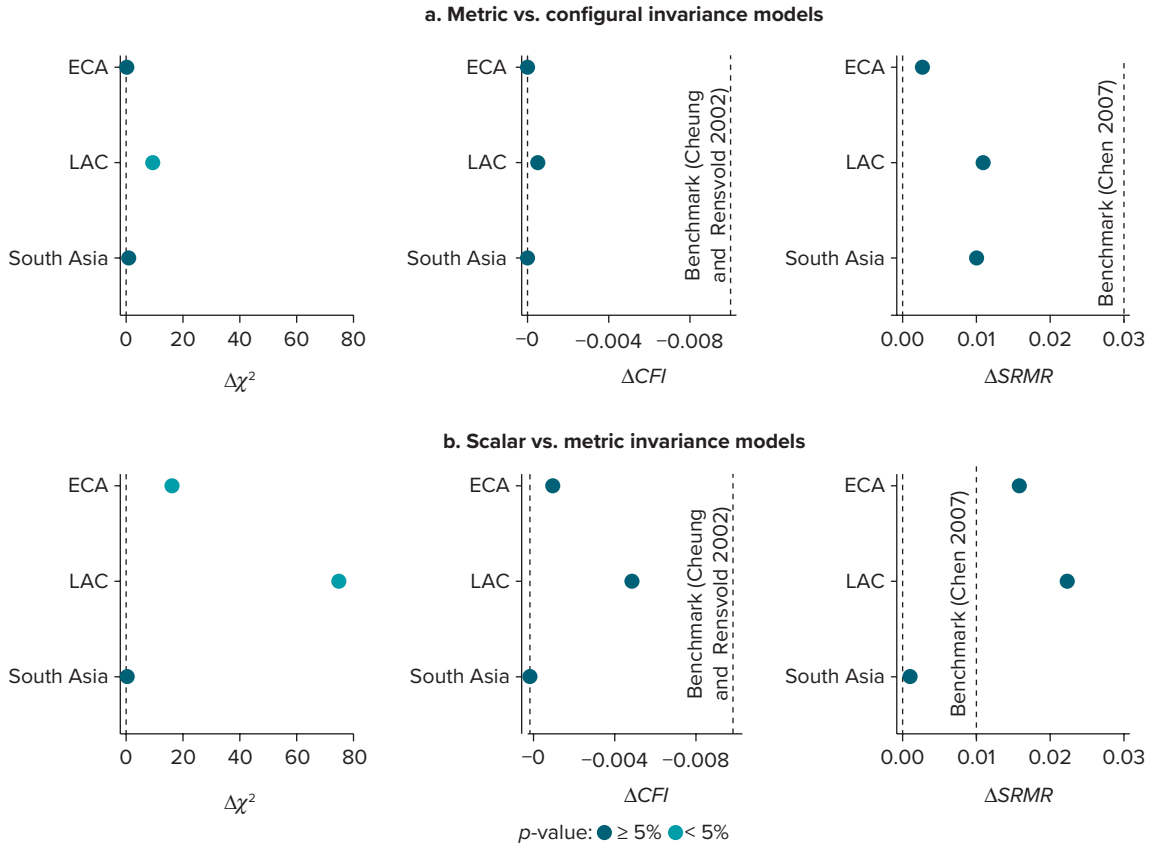
RESULTS

Measurement invariance is tested first in the cross-country context before moving to within-country invariance across demographic groups (gender and education level) and public administration organizations. We start by fitting the cross-country comparison in groupings of countries based on their income and region before moving to compare individual countries to each other. In each case, configural-, metric-, and scalar-invariance models are tested sequentially, provided that the acceptable fit of a higher-level model is first confirmed.

Cross-Country Comparison

The analysis begins with models comparing groups of like countries against each other. It is expected that civil servants in similar countries—that is, those at comparable levels of development or in a single geographical region—are more likely to conceive of transformational leadership in the same manner. Such grouping of countries ensures that the inevitable cultural and socioeconomic differences between countries are minimized. By contrast, comparing a high-income European country, like Estonia, and a large, upper-middle-income country in the heart of Latin America, like Brazil, is a much more demanding test of the invariance concept. Therefore, we move to the latter only after establishing that invariance holds within broader groupings of like countries.

FIGURE 24.2 Measurement Invariance across Countries Classified by Region: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.
Note: Regional classifications are based on World Bank data. The Europe and Central Asia (ECA) Region includes Albania, Kosovo, and Estonia; Latin America and the Caribbean (LAC) includes Brazil and Chile; and South Asia includes Bangladesh and Nepal. CFI = comparative fit index; SRMR = standardized root mean squared error.

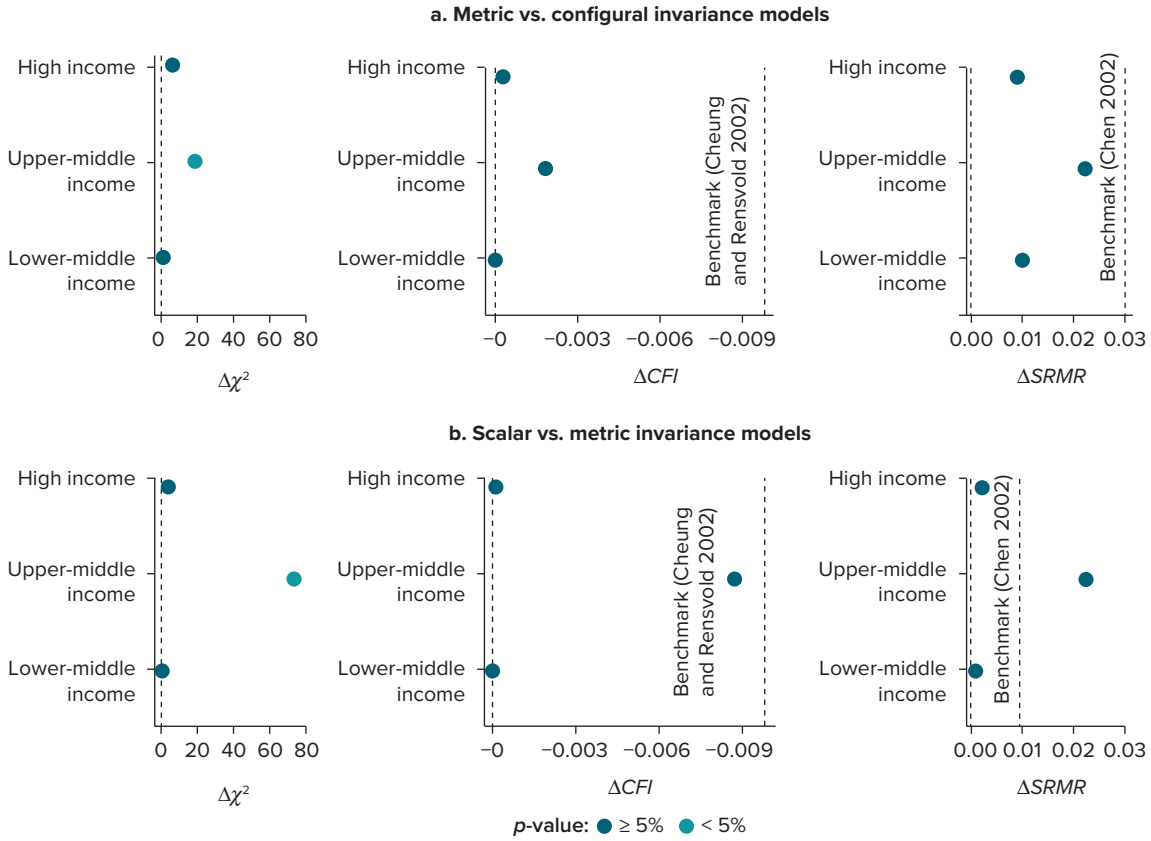
The results of comparing measurement invariance across countries within the same geographical region are shown in figure 24.2. Since all configural-invariance models with only three manifest variables and one latent factor have, by definition, a perfect level of fit, what the figure shows is the change (Δ) in key fit statistics—χ^2, CFI, and SRMR—between configural- and metric-invariance models and between metric- and scalar-invariance ones. For consistency, the changes in the CFI are reversed, meaning that model fit is deteriorating when going right along the *x* axis.

From the figure, it can be seen that metric invariance models fitted across countries from the Europe and Central Asia region (Albania, Estonia, and Kosovo) and for the South Asia region (Bangladesh and Nepal) do not exhibit significantly worse fit on all three indexes. For Latin America and the Caribbean (LAC) (Brazil and Chile), only the $\Delta\chi^2$ is statistically significant, but, given very low changes in the other two indexes, metric invariance can still be inferred.

Taking the metric-invariant models to the next level and imposing scalar invariance, model fit remains fully acceptable for South Asia. For ECA and LAC, both the $\Delta\chi^2$ and the $\Delta SRMR$ point to a significantly worse fit, and, therefore, as with the full cross-country model, scalar invariance can be only tentatively inferred based on the fact that the $\Delta CFI < 0.01$.

Figure 24.3 shows the results of the same analyses replicated across income groupings rather than regions. On the basis of the ΔCFI and the $\Delta SRMR$, all three income groupings exhibit metric invariance. Only a high *p*-value for the upper-middle-income group (Albania, Brazil, and Kosovo) points toward a

FIGURE 24.3 Measurement Invariance across Countries Classified by Income Group: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.
Note: Income groupings are based on World Bank data. High-income countries include Chile and Estonia; upper-middle-income countries include Albania, Brazil, and Kosovo; and lower-middle-income countries include Bangladesh and Nepal. CFI = comparative fit index; SRMR = standardized root mean squared error.

different conclusion, but, as explained in the methodology section, this is not taken as sufficient evidence to overrule a good fit based on the CFI and the SRMR.

If such an interpretation is adopted, scalar-invariance models can be estimated for all income groupings. According to all fit indexes, scalar invariance can be inferred across high-income (Chile and Estonia) and lower-middle-income (Bangladesh and Nepal) countries. For the upper-middle-income countries, the $\Delta\chi^2$ is statistically significant, and the $\Delta SRMR$ is well above the threshold of 0.01. The absolute value of the SRMR of the scalar-invariance model is also only borderline acceptable, at 0.046. The ΔCFI, standing just below 0.009, similarly approaches the threshold of significant deterioration. Therefore, a conclusion of scalar invariance can be drawn only on the basis of the CFI, and, even then, it is not strong. (It should also be noted that the results and conclusions for countries in the lower-middle-income category are exactly the same as for the South Asia category above because those two groups happen to contain the same pair of countries: Bangladesh and Nepal.)

Given the relatively robust evidence of metric and scalar invariance within groupings of comparable countries, we now move to compare all seven countries against each other. When the metric-invariance model is fitted by restricting the factor loadings to be equal across all seven countries, the absolute model fit is still good according to all three fit indexes. The value of χ^2 is 47.1 ($df = 12$), and the associated p-value is close to 0. The CFI drops to 0.998, and the SRMR increases to 0.019. Therefore, the change in the latter two fit indexes is well within the limits recommended by the literature. Although the $\Delta\chi^2$ with a p-value below 5 percent points toward significantly worse fit, the large sample size and perfect fit of the unrestricted model

make this a less reliable measure. Therefore, metric invariance can be inferred for cross-country comparisons of transformational leadership.

Given this conclusion, a scalar-invariance model can be fitted. It represents a borderline case of significant deterioration. The p-value of the $\Delta\chi^2$ is close to 0, and the $\Delta SRMR$ is 0.018, which is above the threshold recommended for scalar-invariance models by Chen (2007). However, the difference is small, and the absolute model fit (the SRMR = 0.037) is still good. Furthermore, the ΔCFI of 0.008 can be viewed as acceptable. Therefore, a tentative conclusion of scalar invariance can be reached.

To summarize the above analyses—in a full cross-country analysis, no significant deterioration in the model of metric invariance suggests that researchers and policy practitioners should be able to compare factor loadings and structural regression coefficients across countries. Item intercepts and means of the indicators can also be compared, although cautiously, given that not all fit indexes suggest that the model with equal intercepts fits the data well.

However, comparisons of this type might be more warranted within groups of like countries. There is evidence that cross-cultural differences in understanding of the idea of transformational leadership are (largely) removed by grouping countries according to their geographical regions. Within such groupings, there is clear evidence of metric invariance. With the same caveat as in the full cross-country models, scalar invariance can also be demonstrated for those models. Invariance is even stronger when countries are grouped by their income level. Both high- and lower-middle-income groups exhibit full metric and scalar invariance. The only group where the conclusion of scalar invariance has very little backing is upper-middle-income countries. This is perhaps unsurprising, given that this group can be viewed as the most heterogeneous, and, therefore, differences in the understanding of concepts such as leadership remain substantial.[5]

Within-Country Comparison: Gender

Turning to intracountry comparisons, gender is the key demographic measure in all public service surveys and also, typically, one of the first lines along which survey results are broken down. The distribution of respondents by gender in the survey sample used here is reported in table 24.3. As can be observed, the gender distribution of civil servants who respond to the surveys varies highly by country. In three out of seven countries, women form the majority of the respondents. The female-to-male ratio varies from approximately 3:1 in Estonia to less than 1:3 in Bangladesh. These cross-country differences are largely consistent with the variation in gender balance across survey populations in countries where personnel records are available (see table 24A.1).

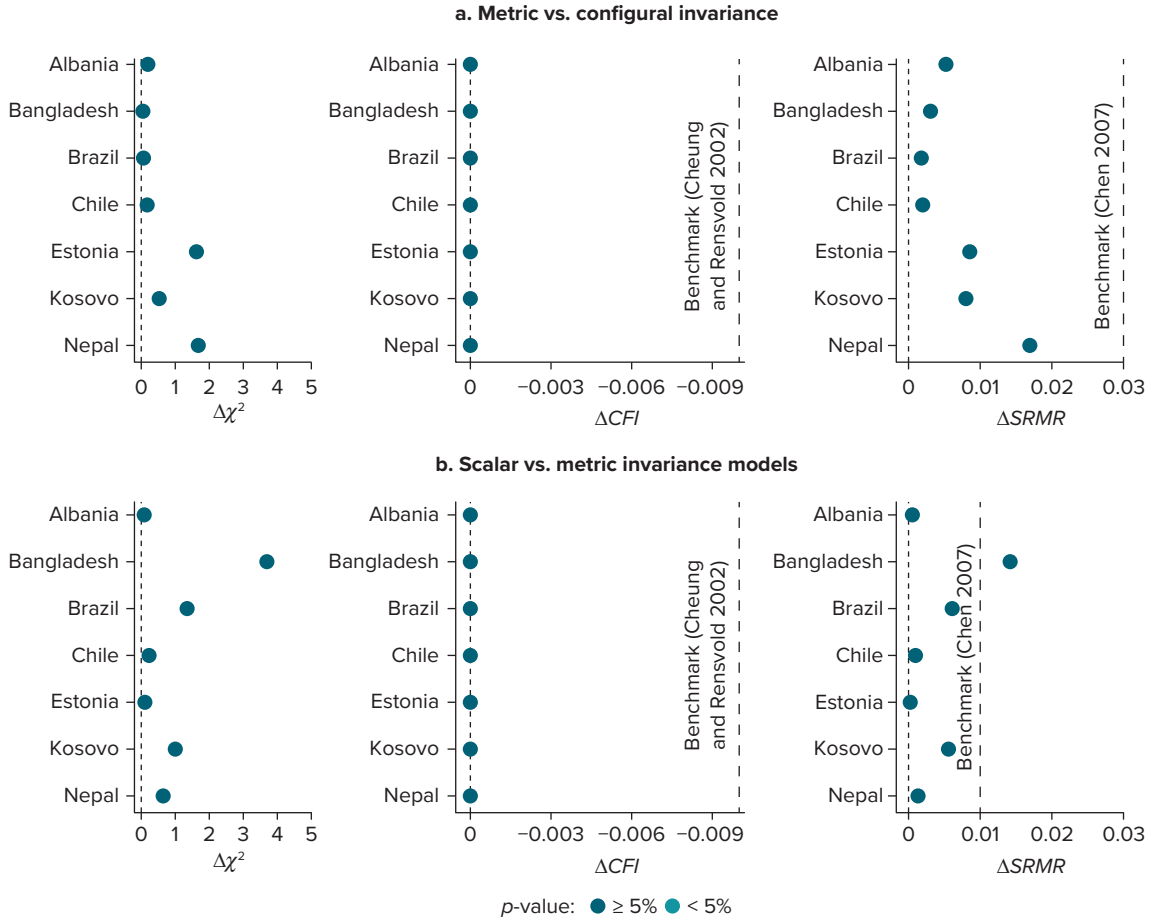
The results of measurement-invariance analyses across gender groups within countries are presented in figure 24.4. Metric invariance—the equality of factor loadings between genders—is obtained with little

TABLE 24.3 Distribution of Respondents, by Gender

Country	Male	Female	Missing
Albania	1,374 (37.2%)	2,261 (61.3%)	55 (1.5%)
Bangladesh	801 (76.4%)	224 (21.4%)	24 (2.3%)
Brazil	2,268 (56.8%)	1,701 (42.6%)	23 (0.6%)
Chile	2,502 (43.6%)	3,155 (54.9%)	85 (1.5%)
Estonia	845 (23.8%)	2,462 (69.3%)	248 (7.0%)
Kosovo	1,363 (55.7%)	1,028 (42.0%)	57 (2.3%)
Nepal	817 (65.4%)	421 (33.7%)	11 (0.9%)

Source: Original table for this publication.

FIGURE 24.4 Measurement Invariance across Gender within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.
Note: CFI = comparative fit index; SRMR = standardized root mean squared error.

space for doubt for all seven countries. In none of them are changes in χ^2 statistically significant, nor are the changes in the CFI and SRMR above their respective thresholds.[6]

As a next step, scalar-invariance models are fitted and compared. Here, the arguments and conclusion remain unchanged. All three fit indexes point to good fit and no significant deterioration of the model after adding equality constraints on item intercepts, which allows us to conclude scalar invariance across genders in all countries considered.

Within-Country Comparison: Education Level

Like with the analyses focused on gender, this subsection concerning education begins with a demographic overview (table 24.4), which presents the distribution of civil servants by their level of education across countries. Here, the heterogeneity across surveys is even more pronounced than in the case of gender. Whereas in Albania, 92.1 percent of civil servants who responded to the survey had university-level education, and only 5.1 percent had below-university-level education, these proportions are equal in Nepal, and in Chile become almost exactly reversed.

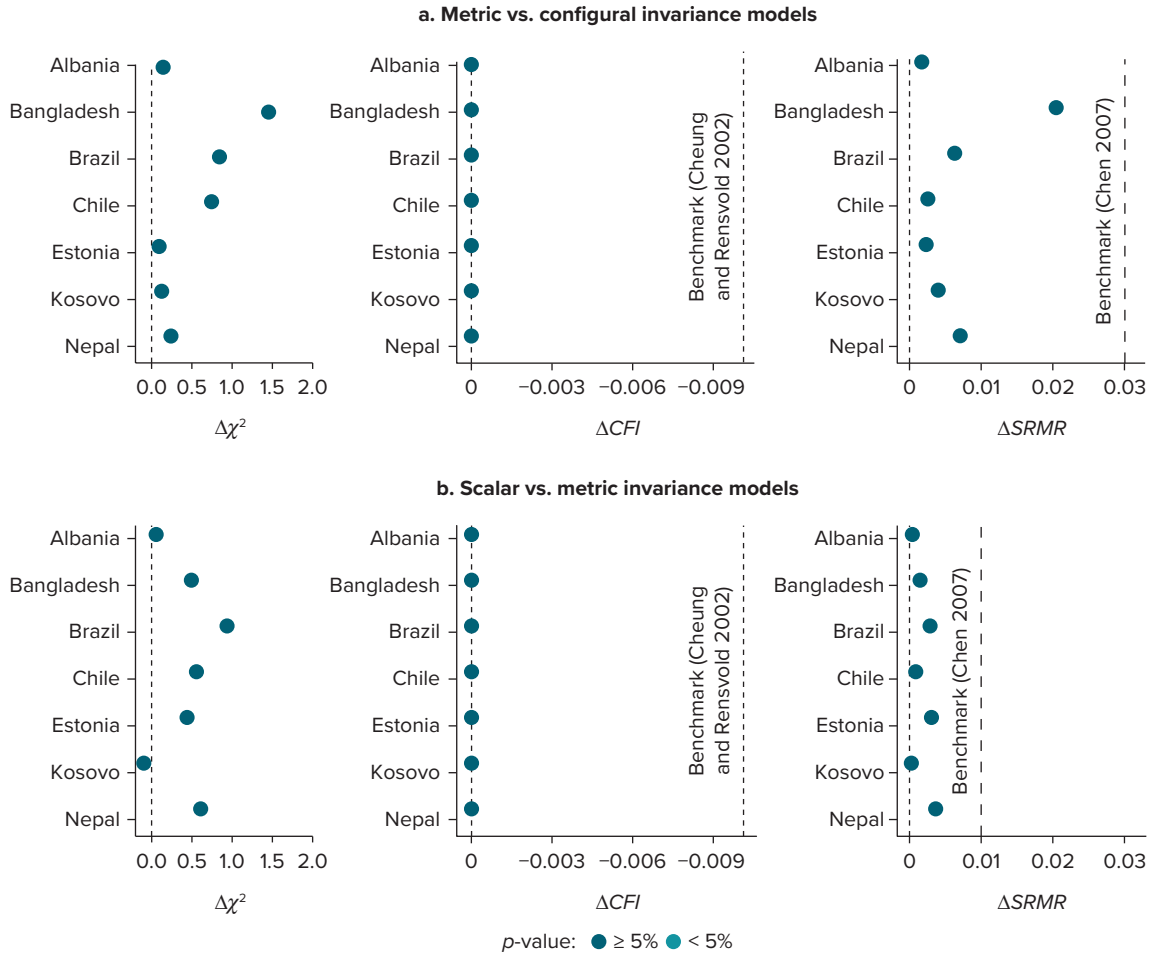
Figure 24.5 demonstrates the results from fitting different levels of invariance models across different education levels in seven countries. As with gender, both metric and scalar invariance can be concluded for all seven countries. None of the changes in the fit indexes come near their respective thresholds.

TABLE 24.4 Distribution of Respondents, by Education Level

Country	University	Below university	Missing
Albania	3,399 (92.1%)	188 (5.1%)	103 (2.8%)
Bangladesh	560 (53.4%)	468 (44.6%)	21 (2.0%)
Brazil	1,964 (49.7%)	1,895 (47.5%)	113 (2.8%)
Chile	586 (10.2%)	5,081 (88.5%)	75 (1.3%)
Estonia	1,898 (53.4%)	1,439 (40.5%)	218 (6.1%)
Kosovo	1,150 (47.0%)	1,261 (51.5%)	37 (1.5%)
Nepal	603 (48.3%)	603 (48.3%)	43 (3.4%)

Source: Original table for this publication.

FIGURE 24.5 Measurement Invariance across Education Levels within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.
Note: CFI = comparative fit index; SRMR = standardized root mean squared error.

Within-Country Comparison: Public Administration Organization

The final set of invariance models is fitted across public administration organizations. The surveys analyzed here were conducted among several central government organizations in each country (see table 24.5). The number of organizations with more than 50 respondents ranges from 7 in Bangladesh to 27 in Estonia. Across countries, the mean number of respondents per organization varies between 83.9 in Kosovo and 522 in Chile. In the latter country, standing at 1,520, the largest number of respondents per organization is also observed.

Figure 24.6 replicates the measurement-invariance comparisons discussed above for gender and education level. However, here the results are less clear-cut. For metric-invariance models, there is clear evidence to suggest the equality of factor loadings for five out of seven countries. For Kosovo and Nepal, the $\Delta SRMR$ is, however, just above 0.03, which suggests significant deterioration compared to the configural-invariance model. Yet the $\Delta\chi^2$ remains small in absolute terms and is also not statistically significant, even though this metric tends to be the most sensitive of the fit indexes. Therefore, metric invariance is concluded for these two countries, albeit with a caveat.

We find similar results when scalar-invariance models are fitted, although here it applies to two additional countries: Bangladesh and Estonia. For these countries, a change in the SRMR points toward significant deterioration in model fit, whereas all other measures suggest acceptable deterioration. Overall, the results suggest that both factor loadings and the means of the transformational-leadership latent factor can be compared across organizations within public administration, but this conclusion is tentative for Kosovo and Nepal, as well as for Bangladesh and Estonia in the case of scalar invariance.

DISCUSSION

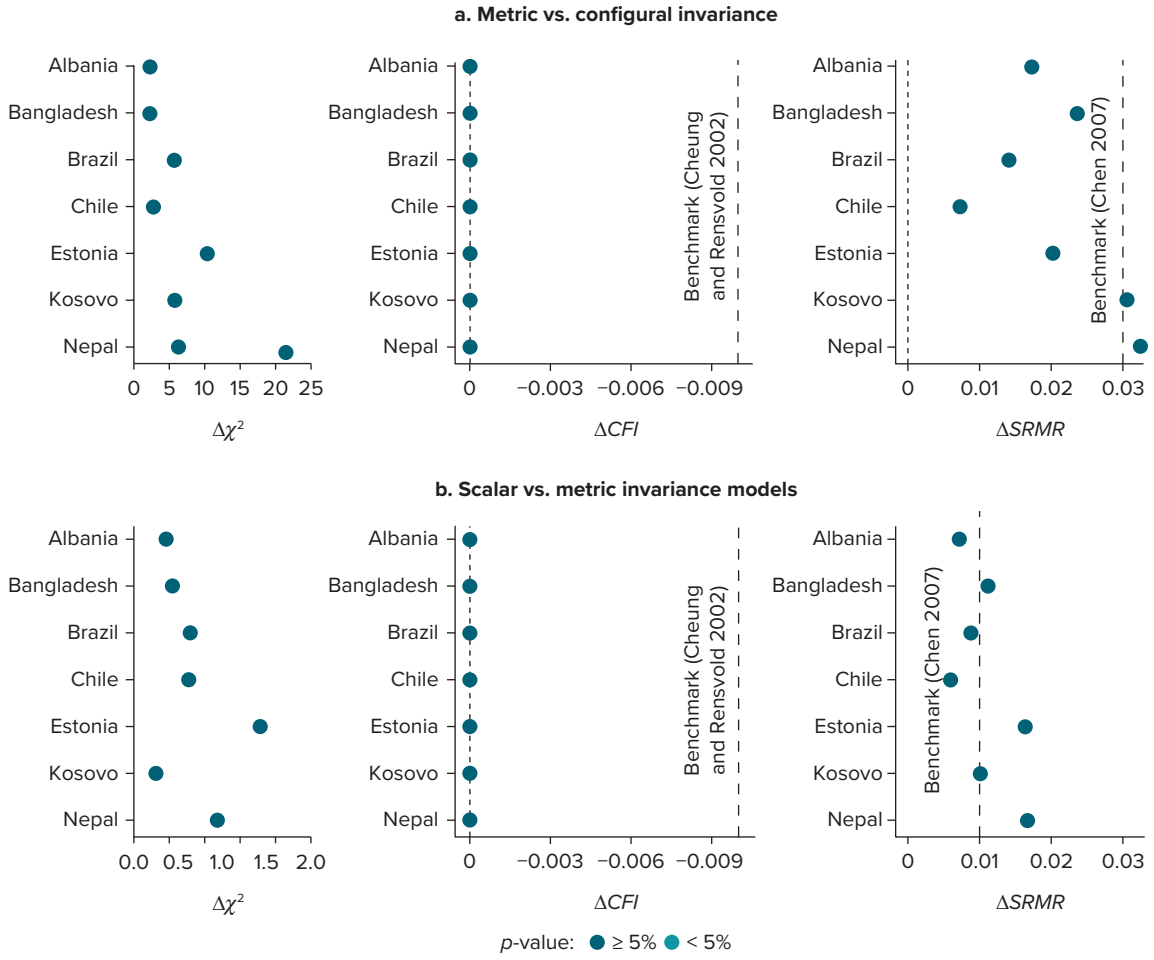
The results of the measurement-invariance analyses of the concept of transformational leadership presented above warrant a tentative two-level conclusion. First, there is strong evidence of metric invariance across countries and tentative evidence of scalar invariance. The latter conclusion can be strengthened if countries are grouped according to region or income level. In that case, full scalar invariance is observed across high-income and South Asian or lower-middle-income countries. Second, transformational leadership appears invariant, both at the level of factor loadings and latent factor means, across gender, broad education level, and organization within public administration in most of the countries studied. The evidence for the

TABLE 24.5 Distribution of Respondents within Public Administration Organizations

Country	No. of respondents					No. of organizations
	Mean	Median	SD	Min.	Max.	
Albania	215.1	183.0	141.9	83	585	15
Bangladesh	120.0	82.0	70.7	52	218	7
Brazil	292.5	165.0	305.9	57	1,062	12
Chile	522.0	382.0	449.4	87	1,520	11
Estonia	108.1	80.0	70.6	64	331	27
Kosovo	83.9	73.5	30.5	54	150	14
Nepal	97.8	83.5	61.1	55	241	8

Source: Original table for this publication.
Note: Only groups with 50+ observations are included in the analyses. SD = standard deviation.

FIGURE 24.6 Measurement Invariance across Public Administration Organizations within Countries: Change in Model Fit Indexes across Configural-, Metric-, and Scalar-Invariance Models



Source: Original figure for this publication.
Note: Only groups with 50+ observations are included in the analyses. CFI = comparative fit index; SRMR = standardized root mean squared error.

first two groups is clear-cut, and, for invariance across organizations, the only caveat that should be raised is the borderline significance of the Δ*SRMR* values in some models.

It is possible that the relatively stronger evidence of invariance within rather than across countries comes from translation differences of the survey items, rather than their differential interpretation. The language of a survey is known to affect the thought and response-forming process of survey respondents, even when care is taken—for instance, through extensive cognitive interviews—to ensure comparable understanding across languages (Peytcheva 2020). Chen (2008) suggests that such difference could also come from a propensity, observable in some cultures, to skew survey responses toward more-neutral options. These possibilities highlight the need not only to standardize the question wording and response scale, as was done here, but also for researchers to retest measurement invariance in public service surveys across further concepts and country settings, including countries with a shared language. Although safeguarding actions were taken to minimize these differences, the results suggest that some residual variation might stem from them.

This unavoidable limitation can, however, serve as a response to another potential criticism of the analyses presented above—namely, the fact that they were focused on only seven countries and a small number of questions concerned with transformational leadership. Including a larger sample of countries would be admittedly desirable, yet it is problematic precisely because the question wording and the wider context of different public service surveys are too dissimilar to warrant inclusion. Most surveys of civil servants include

a leadership section, but they are not directly focused on transformational leadership, and in the cases when they are, their phrasing or response scales make comparisons difficult (chapter 18). Despite including only seven countries, this chapter is, to the best of the authors' knowledge, the first to look at the measurement invariance of the concept of transformational leadership in the public sector in a cross-country context.

CONCLUSION

Surveying civil servants is often the only feasible way to learn more about their attitudes, behaviors, and work environment. Yet survey results are challenging to interpret without context—comparisons to other countries or previous surveys, or between different demographic groups within the public service. However, researchers and policy practitioners only occasionally pause to statistically assess whether the attitudes and behaviors they want to measure—be they engagement, motivation, or leadership—are understood in the same way by different groups of civil servants.

Drawing on the concept of measurement invariance, this chapter was able to show that when it comes to the differential understanding of the concept of transformational leadership, differences in gender, education level, and organization have a very small impact. In most countries, the three leadership questions relate in the same way to the underlying concept of transformational leadership, and the mean levels of transformational leadership are also comparable across groups. The same can be said when looking across countries, even if the conclusion of the comparability of the mean levels is weakened, to an extent, when comparing countries at different income levels and, in particular, in different regions. This suggests, tentatively, that global benchmarking exercises like the GSPS have a legitimate empirical foundation. Contrasting our results with those of measurement-invariance analyses of PSM—a concept which is arguably even more culturally specific than leadership—suggests, in particular, that questions that are less culturally loaded and more factual—for instance, about management practices rather than culturally specific attitudes—might have a stronger empirical basis for comparison.

Of course, we assess only one measurement scale in our analysis and draw on data from seven countries. Thus, much fertile empirical ground for future cross-country work on measurement invariance remains to further solidify claims about what can be compared across countries and what cannot. At least four further contributions would be especially welcome in the future. First, future investigations should extend analyses of measurement invariance to other recurring topics in public service surveys. Perceptions about work environment, engagement, teamwork, compensation, turnover, performance, meritocratic practices, and harassment are components of many public service surveys. Yet the extent to which they measure the same underlying concepts across different groups of civil servants and across countries is uncertain. A second possible extension of the present analyses would include more countries in the analyses, preferably with heterogeneous geographical and economic features. A third avenue for future work would address the fact that at present, only a limited number of groupings of civil servants have been compared for measurement invariance. This chapter focused on gender, education level, and organization. Including further groupings—by age and tenure level, managerial position, and contract type—would be warranted in future studies. A fourth type of analysis would ascertain intertemporal measurement invariance. Just as the same question can measure divergent concepts across different countries, cultures, or demographic groups, it can be measurement variant across different time periods. Due to changes in social, economic, political, and, in the longer term, cultural conditions, the same survey question might come to be interpreted differently in different time periods, even when asked to the same population.

Along with further investigations of measurement invariance, researchers and practitioners wishing to compare the results of surveys of public servants would be well served by relying, at least in part, on a standardized questionnaire. One such effort is the GSPS initiative, which catalogs 20+ sets of public service survey results, along with their respective questionnaires, section names, and metadata. Including some of the standardized questions would allow for survey results to be more readily compared with other countries'

results and, ultimately, for the establishment of international benchmarks against which civil servants' attitudes and behaviors could be reliably compared. Even when such comparisons are tentative, given concerns with measurement invariance, this certainly trumps comparisons of core concepts (for example, employee engagement) across countries using different measures.

NOTES

We are grateful to Daniel Rogger and Galileu Kim for helpful comments.

1. Unless justified for other reasons, in all instances countries are listed in alphabetical order. See the GSPS website (https://www.globalsurveyofpublicservants.org/) and Mikkelsen, Schuster, and Meyer-Sahling (2020) for further details on the surveys included.
2. Technical limitations, like limited access to electricity, computer, or the Internet, as well as incomplete databases of email records for civil service officials, made online surveying unfeasible in the two Asian countries (Bangladesh and Nepal) included in this chapter.
3. State or local government officials and nonadministrative public sector employees, like teachers, nurses, doctors, policemen, and the military, were thus excluded.
4. It was decided that another commonly employed measure of model fit, the root mean square error of approximation (RMSEA), would not be used in the analyses presented here. The RMSEA was introduced by Steiger and Lind (1980) and extended by, among others, Browne and Cudeck (1993) and Steiger (1998). However, it can be unreliable when comparing just-identified with overidentified models, as is done here. Using Monte Carlo simulations, Kenny, Kaniskan, and McCoach (2015) find that in models with few degrees of freedom, the RMSEA tends to be overinflated and, therefore, falsely points to bad model fit. Moreover, in close-fit models, more restricted models might counterintuitively show a decrease in the RMSEA—that is, better fit—because of the increased number of degrees of freedom (Shi, Lee, and Maydeu-Olivares 2019). Notwithstanding the above, RMSEA values point toward the same broad conclusions as the other three fit indexes consulted in the text.
5. In contrast, the lower-middle-income group is relatively homogeneous, since it is comprised of two South Asian countries.
6. In fact, it can be observed the ΔCFI is 0 across all intracountry comparisons. This is because the model fit is close to perfect, and, as a result, χ^2 is low enough as to be smaller than the number of degrees of freedom. Given the formula used to calculate the CFI, the resulting value of this fit index will always be 1 in those cases (see Bentler 1990).

REFERENCES

Bass, B. M. 1985. *Leadership and Performance beyond Expectations*. New York: Free Press.

Bentler, P. 1990. "Comparative Fit Indices in Structural Models." *Psychological Bulletin* 107 (2): 238–46. https://doi.org/10.1037/0033-2909.107.2.238.

Bentler, P. 1995. *EQS 5* [Computer program]. Encino, CA: Multivariate Software.

Berson, Y., and B. J. Avolio. 2004. "Transformational Leadership and the Dissemination of Organizational Goals: A Case Study of a Telecommunication Firm." *The Leadership Quarterly* 15 (5): 625–46. https://doi.org/1016/j.leaqua.2004.07.003.

Browne, M. W., and R. Cudeck. 1993. "Alternative Ways of Assessing Model Fit." In *Testing Structural Equation Models*, edited by K. A. Bollen and J. S. Long, 136–62. Newbury Park, CA: Sage.

Burns, J. M. 1978. *Leadership*. New York: Harper & Row.

Byrne, B. M., R. H. Shavelson, and B. Muthén. 1989. "Testing for the Equivalence of Factor Covariance and Mean Structures: The Issue of Partial Measurement Invariance." *Psychological Bulletin* 105 (3): 456–66. https://doi.org/10.1037/0033-2909.105.3.456.

Cabinet Office. 2019. *Civil Service People Survey 2019: Technical Guide*. London: Cabinet Office, United Kingdom Government. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachmentdata/file/867302/Civil-Service-People-Survey-2019-Technical-Guide.pdf.

Chen, F. F. 2007. "Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 14 (3): 464–504. https://doi.org/10.1080/10705510701301834.

Chen, F. F. 2008. "What Happens If We Compare Chopsticks with Forks? The Impact of Making Inappropriate Comparisons in Cross-Cultural Research." *Journal of Personality and Social Psychology* 95 (5): 1005–18. https://doi.org/10.1037/a0013193.

Cheung, G. W., and R. B. Rensvold. 2002. "Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 9 (2): 233–55. https://doi.org/10.1207/S15328007SEM0902_5.

Cieciuch, J., E. Davidov, P. Schmidt, and R. Algesheimer. 2019. "How to Obtain Comparable Measures for Cross-National Comparisons." *Kolner Zeitschrift fu¨r Soziologie und Sozialpsychologie* 71 (S1): 157–86. https://doi.org/10.1007/s11577-019-00598-7.

Davidov, E., H. Dülmer, J. Cieciuch, A. Kuntzm, D. Seddig, and P. Schmidt. 2018. "Explaining Measurement Nonequivalence Using Multilevel Structural Equation Modeling: The Case of Attitudes toward Citizenship Rights." *Sociological Methods & Research* 47 (4): 729–60. https://doi.org/10.1177/0049124116672678.

Davidov, E., B. Meuleman, J. Cieciuch, P. Schmidt, and J. Billiet. 2014. "Measurement Equivalence in Cross-National Research." *Annual Review of Sociology* 40 (1): 55–75. https://doi.org/10.1146/annurev-soc-071913-043137.

De Jong, M. G., J.-B. Steenkamp, and J.-P. Fox. 2007. "Relaxing Measurement Invariance in Cross-National Consumer Research Using a Hierarchical IRT Model." *Journal of Consumer Research* 34 (2): 260–78. https://doi.org/10.1086/518532.

Donkor, F., I. Sekyere, and F. A. Oduro. 2022. "Transformational and Transactional Leadership Styles and Employee Performance in Public Sector Organizations in Africa: A Comprehensive Analysis in Ghana." *Journal of African Business* 23 (4): 945–63. https://doi.org/10.1080/15228916.2021.1969191.

Downton, J. V. 1973. *Rebel Leadership: Commitment and Charisma in the Revolutionary Process*. New York: Free Press.

Erkutlu, H. 2008. "The Impact of Transformational Leadership on Organizational and Leadership Effectiveness: The Turkish Case." *Journal of Management Development* 27 (7): 708–26. https://doi.org/10.1108/02621710810883616.

Fitzpatrick, J., M. Goggin, T. Heikkila, D. Klingner, J. Machado, and C. Martell. 2011. "A New Look at Comparative Public Administration: Trends in Research and an Agenda for the Future." *Public Administration Review* 71 (6): 821–30. https://doi.org/10.1111/j.1540-6210.2011.02432.x.

French, B. F., and H. W. Finch. 2006. "Confirmatory Factor Analytic Procedures for the Determination of Measurement Invariance." *Structural Equation Modeling* 13 (3): 378–402. https://doi.org/10.1207/s15328007sem1303_3.

Fukuyama, F., D. Rogger, Z. Hasnain, K. Bersch, D. Mistree, C. Schuster, K. Mikkelsen, K. Kay, and J. Meyer-Sahling. 2022. *Global Survey of Public Servants Indicators*. https://www.globalsurveyofpublicservants.org.

García-Morales, V. J., F. J. Lloréns-Montes, and A. J. Verdú Jover. 2008. "The Effects of Transformational Leadership on Organizational Performance through Knowledge and Innovation." *British Journal of Management* 19 (4): 299–319. https://doi.org/10.1111/j.1467-8551.2007.00547.x.

Hameduddin, T., and T. Engbers. 2021. "Leadership and Public Service Motivation: A Systematic Synthesis." *International Public Management Journal* 25 (1): 86–119. https://doi.org/10.1080/10967494.2021.1884150.

Hofman, D. A., J. E. Mathieu, and R. Jacobs. 1990. "A Multiple Group Confirmatory Factor Analysis Evaluation of Teachers' Work Related Perceptions and Reactions." *Educational and Psychological Measurement* 50 (4): 943–55. https://doi.org/10.1177/0013164490504024.

Hong, S., M. L. Malik, and M.-K. Lee. 2003. "Testing Configural, Metric, Scalar, and Latent Mean Invariance across Genders in Sociotropy and Autonomy Using a Non-Western Sample." *Educational and Psychological Measurement* 63 (4): 636–54. https://doi.org/10.1177/0013164403251332.

Hooper, D., J. Coughlan, and M. R. Mullen. 2008. "Structural Equation Modelling: Guidelines for Determining Model Fit." *The Electronic Journal of Business Research Methods* 6 (1): 53–60.

Hu, L.-T., and P. 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. https://doi.org/10.1080/10705519909540118.

Jensen, U. T., L. B. Andersen, L. L. Bro, A. Bøllingtoft, T. L. Eriksen, A.-L. Holten, C. B. Jacobsen, et al. 2019. "Article Conceptualizing and Measuring Transformational and Transactional Leadership." *Administration & Society* 51 (1): 3–33. https://doi.org/10.1177/0095399716667157.

Jreisat, J. G. 2005. "Comparative Public Administration Is Back in, Prudently." *Public Administration Review* 65 (2): 231–42. https://doi.org/10.1111/j.1540-6210.2005.00447.x.

Kenny, D. A., B. Kaniskan, and B. D. McCoach. 2015. "The Performance of RMSEA in Models with Small Degrees of Freedom." *Sociological Methods & Research* 44 (3): 486–507. https://doi.org/10.1177/0049124114543236.

Kim, S., W. Vandenabeele, B. E. Wright, L. B. Andersen, F. P. Cerase, R. K. Christensen, C. Desmarais, 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. https://doi.org/10.1093/jopart/mus027.

Kroll, A., and D. Vogel. 2014. "The PSM–Leadership Fit: A Model of Performance Information Use." *Public Administration* 92 (4): 974–91. https://doi.org/10.1111/padm.12014.

La Salle, T. P., D. B. McCoach, and J. Meyers. 2021. "Examining Measurement Invariance and Perceptions of School Climate across Gender and Race and Ethnicity." *Journal of Psychoeducational Assessment* 39 (7): 800–15. https://doi.org/10.1177/07342829211023717.

Li, C.-H. 2016. "Confirmatory Factor Analysis with Ordinal Data: Comparing Robust Maximum Likelihood and Diagonally Weighted Least Squares." *Behavioral Research Methods* 8 (3): 936–49. https://doi.org/10.3758/s13428-015-0619-7.

Martinez, A. J. 2021. "Factor Structure and Measurement Invariance of the Academic Time Management and Procrastination Measure." *Journal of Psychoeducational Assessment* 39 (7): 891–901. https://doi.org/10.1177/07342829211034252.

Meredith, W. 1964. "Notes on Factorial Invariance." *Psychometrika* 29: 177–85. https://doi.org/10.1007/BF02289699.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, C. Pesti, and T. Randma-Liiv. 2018a. *Civil Service Management in Estonia: Evidence from a Survey of Civil Servants and Employees*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. Last updated October 2018. https://christianschuster.net/Meyer-Sahling%20Schuster%20Mikkelsen%20Pesti%20Randma-Liiv%20Estonia%20Report%20FINAL.pdf.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, H. Qeriqi, and F. Toth. 2018b. *Towards a More Professional Civil Service in Kosovo: Evidence from a Survey of Civil Servants in Central and Local Government*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. https://christianschuster.net/Meyer-Sahling%20Schuster%20Mikkelsen%20Qeriqi%20Toth%20Kosovo%20Report%20FINAL.pdf.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, T. Rahman, K. M. Islam, A. S. Huque, and F. Toth. 2019. *Civil Service Management in Bangladesh: Evidence from a Survey of More Than 1,000 Civil Servants*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. https://christianschuster.net/2019.03.10.%20Bangladesh%20FOR%20PUB LICATION.pdf.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, S. K. Shrestha, B. Luitel, and F. Toth. 2018c. *Civil Service Management in Nepal: Evidence from a Survey of More than 1,200 Civil Servants*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. https://christianschuster.net/2018.12.20.%20Nepal%20FOR%20PUBLICATION.pdf.

Meyer-Sahling, J., C. Schuster, K. Sass Mikkelsen, and A. Shundi. 2018d. *The Quality of Civil Service Management in Albania: Evidence from a Survey of Central Government Civil Servants and Public Employees*. London: British Academy, UK Department for International Development Anti-Corruption Evidence Programme. https://christianschuster.net/Meyer-Sahling%20Schuster%20Mikkelsen%20Shundi%20Albania%20Report%20FINAL.pdf.

Mikkelsen, K. S., C. Schuster, and J.-H. Meyer-Sahling. 2020. "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." *International Public Management Journal* 24 (6): 739–61. https://doi.org/10.1080/10967494.2020.1809580.

Nguyen, T. T., L. Mia, L. Winata, and V. K. Chong. 2017. "Effect of Transformational-Leadership Style and Management Control System on Managerial Performance." *Journal of Business Research* 70: 202–31. https://doi.org/10.1016/j.jbusres.2016.08.018.

OPM (Office of Personnel Management). 2019. *2019 Office of Personnel Management Federal Employee Viewpoint Survey: Technical Report*. Washington, DC: US Office of Personnel Management, US Government.

Pandey, S. K., R. S. Davis, S. Pandey, and S. Peng. 2016. "Transformational Leadership and the Use of Normative Public Values: Can Employees Be Inspired to Serve Larger Public Purposes?" *Public Administration* 94 (1): 204–22. https://doi.org/10.1111/padm.12214.

Park, S. M., and H. G. Rainey. 2008. "Leadership and Public Service Motivation in U.S. Federal Agencies." *International Public Management Journal* 11 (1): 109–42. https://doi.org/10.1080/10967490801887954.

Pearce, C. L., H. P. Sims Jr., J. F. Cox, G. Ball, E. Schnell, K. A. Smith, and L. Trevino. 2002. "Transactors, Transformers and Beyond. A Multi-Method Development of a Theoretical Typology of Leadership." *Journal of Management Development* 22 (4): 273–307. https://doi.org/10.1108/02621710310467587.

Pereira, A. K., R. A. Machado, P. L. Costa Cavalcante, A. De Avila Gomide, A. Gomes Magalhaes, I. De Araujo Goellner, R. R. Coelho Pires, K. Bersch, F. Fukuyama, and A. R. Da Silva. 2021. "Government Quality and State Capacity: Survey Results from Brazil." CDDRL Working Paper, Center on Democracy, Development, and the Rule of Law, Stanford University, Stanford, CA. https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/governancereportbrazil.pdf.

Peytcheva, E. 2020. "The Effect of Language of Survey Administration on the Response Formation Process." In *The Essential Role of Language in Survey Research*, edited by M. Sha and T. Gabel, 3–22. Research Triangle Park, NC: RTI Press.

Putnick, D. L., and M. H. Bornstein. 2016. "Measurement Invariance Conventions and Reporting: The State of the Art and Future Directions for Psychological Research." *Developmental Review* 41: 71–90. https://doi.org/10.1016/j.dr.2016.06.004.

Rosseel, Y. 2012. "Lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software* 48 (2): 1–36. https://doi.org/10.18637/jss.v048.i02.

Schuster, C., J. Fuenzalida, J.-H. Meyer-Sahling, K. Mikkelsen, and N. Titelman. 2020. *Encuesta Nacional de Funcionarios en Chile: Evidencia para un servicio público más motivado, satisfecho, comprometido y ético* [National Survey of Civil Servants in Chile: Evidence for a More Motivated, Satisfied, Engaged, and Ethical Public Service]. Santiago: Dirección Nacional del

Servicio Civil. https://www.serviciocivil.cl/wp-content/uploads/2020/01/Encuesta-Nacional-de-Funcionarios-Informe-General-FINAL-15ene2020-1.pdf.

Schuster, C., J. Meyer-Sahling, K. S. Mikkelsen, and C. González Parrao. 2017. *Prácticas de gestión de personas para un servicio público más motivado, comprometido y ético en Chile: Evidencia de una encuesta con 20.000 servidores públicos en Chile y otros países*. Santiago: Dirección Nacional del Servicio Civil. https://documentos.serviciocivil.cl/actas/dnsc/documentService/downloadWs?uuid=60fcd3de-fa9e-4906-9396-c7637b4cd167%20.

Seddig, D., and H. Leitgöb. 2018. "Approximate Measurement Invariance and Longitudinal Confirmatory Factor Analysis: Concept and Application with Panel Data." *Survey Research Methods* 12 (1): 29–41.

Shi, D., T. Lee, and A. Maydeu-Olivares. 2019. "Understanding the Model Size Effect on SEM Fit Indices." *Educational and Psychological Measurement* 79 (2): 310–34. https://doi.org/10.1177/0013164418783530.

Shi, D., A. Maydeu-Olivares, and Y. Rosseel. 2020. "Assessing Fit in Ordinal Factor Analysis Models: SRMR vs. RMSEA." *Structural Equation Modeling: A Multidisciplinary Journal* 27 (1): 1–15.

Steenkamp, J.-B., and H. Baumgartner. 1998. "Assessing Measurement Invariance in Cross-National Consumer Research." *Journal of Consumer Research* 25 (1): 78–90. https://doi.org/10.1086/209528.

Steiger, J. H. 1998. "A Note on Multiple Sample Extensions of the RMSEA Fit Index." *Structural Equation Modelling* 5 (4): 411–19. https://doi.org/10.1080/10705519809540115.

Steiger, J. H., and J. C. Lind. 1980. "Statistically Based Tests for the Number of Common Factors." Paper presented at the annual Spring Meeting of the Psychometric Society, Iowa City, IA.

Struening, E. L., and J. Cohen. 1963. "Factorial Invariance and Other Psychometric Characteristics of Five Opinions about Mental Illness Factors." *Educational and Psychological Measurement* 23: 289–98. https://doi.org/10.1177/001316446302300206.

Tummers, L., and E. Knies. 2016. "Measuring Public Leadership: Developing Scales for Four Key Public Leadership Roles." *Public Administration* 94 (2): 433–51. https://doi.org/10.1111/padm.12224.

Vandenberg, R. J., and C. E. Lance. 2000. "A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research." *Organizational Research Methods* 3 (1): 4–70. https://doi.org/10.1177/109442810031002.

Van De Vijveri, F. J. R., F. Avvisatiii, E. Davidoviii, M. Eidiv, J.-P. Foxv, N. Le Donnéii, K. Lekvi, B. Meulemanvii, M. Paccagnellaii, and R. Van De Schoot. 2019. "Invariance Analyses in Large-Scale Studies." OECD Education Working Paper 201, OECD, Paris. https://doi.org/10.1787/254738dd-en.

Wang, G., I.-S. Oh, S. H. Courtright, and A. E. Colbert. 2011. "Transformational Leadership and Performance across Criteria and Levels: A Meta-Analytic Review of 25 Years of Research." *Group & Organization Management* 36 (2): 223–70. https://doi.org/10.1177/1059601111401017.