CHAPTER 59

Challenges for Large-Scale International Comparative Survey-Based Research in Public Administration

Koen Verhoest, Jan Wynen, Wouter Vandenabeele and Steven Van de Walle

Abstract Within the field of public administration, the experience with large-scale international comparative survey-based research is expanding. However, while such research might render extremely interesting data for comparison and analysis, such resource and time-intensive research bring also challenges, in terms of methodological risks, data quality issues, and governance. This chapter uses the experience of the COBRA survey, the COCOPS survey, and the International Public Service Motivation survey and discusses methodological challenges, like the non-response bias, the issue of measurement invariance, common method bias, and limitations of cross-sectional data, as well

K. Verhoest (✉) · J. Wynen
Research Group on Public Administration and Management,
Department of Political Science, University of Antwerp, Antwerp, Belgium
e-mail: koen.verhoest@uantwerpen.be

J. Wynen
e-mail: jan.wynen@uantwerpen.be

W. Vandenabeele
Utrecht University School of Governance, Utrecht, The Netherlands
e-mail: w.v.vandenabeele@uu.nl

J. Wynen · S. Van de Walle
KU Leuven Public Governance Institute, KU Leuven, Belgium
e-mail: steven.vandewalle@kuleuven.be

E. Ongaro and S. van Thiel (eds.), *The Palgrave Handbook of Public Administration and Management in Europe*,
https://doi.org/10.1057/978-1-137-55269-3_59

as important success factors for the governance of such projects. This chapter ends with some reflections on potential strategies of the public administration community to further enhance and support such research, also toward funders and governments.

59.1 Introduction: Where Do We Stand?[1]

Country-specific politico-administrative regimes, institutions, and administrative traditions matter in how public administration and management develop and functions. Hence, among public administration scholars, there is a wide-held agreement that in public administration internationally comparative research is necessary in the quest for knowledge about general causal mechanisms and about the influence of such institutions and traditions. Cross-country comparative research, using quantitative and qualitative methods, is certainly gaining importance in our field.

This chapter focuses on international comparative research through large-scale surveys, in which similar phenomena are studied by replicating the same questionnaire in a substantial number of countries. Although this kind of research is not new in other domains of political science, and also not in public administration (e.g., the UDITE survey on CEOs in local government in 13 countries during 1995–1997), the experience of the European public administration community is rather limited. In the last decade some significant initiatives have been taken (see below), but surely the opportunities for such large-scale survey-based international comparative research are increasing. Research and network funding programs from the EU provide possibilities to get sufficient financial resources for such research. Moreover, European and national policy-makers increasingly see the merit for international comparative public administration research and practitioners from the European civil services value insights coming from such research.

However, how do we make sure that we get maximum value from such large-scale survey-based comparative research? It is clear that setting-up, conducting, and managing such research is resource and time-intensive. How can we make sure such coordinated research efforts are worth the investments they take in terms of resources, coordination efforts, and time? It is not uncommon that the data collection phase consumes most of the available resources of the participating research groups, with under-exploitation of the resulting data and decreasing involvement of partners in the data-analysis phase as a consequence.

The question whether and how such large-scale survey-based comparative research can have maximum value becomes even more pertinent, considering the increasing focus of high-ranked journals on potential biases in survey-based research when reviewing and accepting publications. We might refer to recent editorials and discussions of potential common method bias/common source bias in studies which rely upon the same survey for the main variables in their study (Kelman 2015; Jakobsen and Jensen 2015). But one can also think of other endogeneity problems like omitted variable bias or reversed causality

claims that reviewers might make when commenting pieces using such comparative surveys. In general, there is an increased skepticism regarding the value of cross-sectional surveys for explanatory analyses. However, the question how to deal with these issues gets another flavor in the context of international comparative research, collecting data in 10–20 countries through surveys.

In this chapter we discuss some of these questions, learning from the experience of three such initiatives, being the COBRA survey on autonomy and control of public agencies (see Verhoest et al. 2012), the COCOPS survey on public sector reforms (Hammerschmid et al. 2016) and the International Public Service Motivation survey (Kim et al. 2013).

These three initiatives are interesting to learn from because they had at least to some extent different objectives as well as different origins, which impacted to a significant extent their methodological choices as well as their governance approach. In the next section, we highlight the differences between the three initiatives in terms of objectives and origins.

Each of these three initiatives replicated a common survey in an international setting confronting them not only with the usual challenges of survey research, but also with additional challenges stemming from the fact that the survey is run in several countries. In the third section of this chapter, we focus on two of these methodological challenges in a more elaborated way: the non-response bias, the issue of measurement invariance, and discuss some potential ways in which we can overcome potential limitations of cross-sectional international surveys.

We can also learn from these experiences about how to organize the governance of such research efforts. In these three initiatives, the number of countries in which the surveys were replicated ranged from 12 to 20 countries, mostly within Europe, but also sometimes outside Europe. These research efforts involved a large number of research groups and individual researchers, complicating the governance of such networks. In the fourth section, we briefly touch upon the most important aspects of governance.

We conclude this chapter by discussing some issues for the public administration community.

59.2 How Has It Been Done Before?

The three initiatives from which we draw lessons are quite different in terms of their origins and objectives, resulting in different methodological and governance choices. In this section, we briefly present these initiatives and their objectives, development, targeted respondents, and scope and end with some reflections.

The COBRA initiative refers in this context to an international survey on public agencies by the "Comparative Public Organization Data Base for Research and Analysis – network",[2] an academic research network which had been initiated at the Public Management Institute of the KULeuven by Geert Bouckaert and Guy Peters with support from Koen Verhoest and Bram Verschuere in 2001. The COBRA network aimed at further developing and replicating a common

questionnaire in order to survey senior managers of (semi-)autonomous public agencies in order to build a cross-country database for comparative, descriptive, and explanatory analysis. The survey has been focusing on autonomy and control of public agencies as well as organizational variables potentially affecting/affected by autonomy and control (Verhoest et al. 2012).

A very particular feature of this initiative is its gradual development and expansion over a period of almost 15 years. The COBRA survey was initially developed, tested, and conducted by Geert Bouckaert, Koen Verhoest, and Bram Verschuere of the Public Management Institute (KULeuven) in collaboration with B. Guy Peters in 2002–2004. The questionnaire was a combination of case-study based new conceptualization and operationalization for some core concepts (like autonomy and control, see Verhoest et al. 2004), together with previously used items and questions for other concepts (like organizational culture). The initial survey was further refined and streamlined within the "Comparative Public Organization Data Base for Research and Analysis" (COBRA) network in a gradual process. First, the original Belgian survey was replicated in 2004 in Norway and Ireland, but adding new questions and experimenting with other answer categories. After this first wave, a consolidated version of the survey was agreed upon within the network, defining three categories of questions, of which the first, 'core' set of questions was to be applied in the same form in all subsequent survey replications in other countries, while the two other categories had more degrees of freedom. From that moment onwards, the participating countries and research groups expanded in three subsequent 'waves', a process which was intensified by integrating this survey in the CRIPO-COST Action, "Comparative Research into Current Trends in Public Sector Organization", funded between 2007 and 2011 by the EU Cost Action program. By the end of the CRIPO-COST Action in 2011 the survey had been replicated in 15 European countries,[3] the EU-level, and three non-European countries (Australia, Hong Kong, Tanzania), gathering data for 1769 organizations and feeding into a joint COBRA database (see for more information about countries, types of agencies surveyed and response rates, Verhoest et al. 2012). Since 2011 there was further survey replication in Spain, Pakistan, and in the German 'Länder'.

The survey was sent in each country to the CEOs of the full population of certain types of agencies. When replicating the survey, each country team had to map the population of agencies first, collect some factual data, translate and pilot-test the core survey, and discuss any issues with the central coordinator, which monitored these processes in order to maintain the comparability between the surveys over time. When translating the survey, special care was taken for taking into account the particularities of the country's administrative system, the position of agencies in these systems and the use of certain country-specific terms. In the survey, factual questions were emphasized in order to reduce interpretation and measurement problems. Throughout the process of expansion, country teams added new questions of which a few became part of the core set of questions replacing some of the less relevant core questions. While the initial focus of the survey on the autonomy and control of agencies in

relation with their minister and parent department remained central, these new questions allowed a more relational view of agency's autonomy by mapping the influence and interactions with other actors, including the European Commission (see Yesilkagit and van Thiel 2012; Bach et al. 2015). So, the expansion of the survey across countries was also a process of evolution and joint learning.

The gradual expansion to new countries was mainly a process of community-building and persuasion, as the research efforts of each of the involved teams had to be funded by their own means or by national research or government funds. The process of getting research teams involved to a large extent helped initially by the reputation of and promotion by the project champions (Geert Bouckaert, Guy Peters, and Per Laegreid) and later by the COST funding for network activities and publications and the visibility of the network. In total, about 50 researchers were involved. A code of conduct was agreed upon regulating data ownership and joint publications. While during the gradual expansion the comparability of data remained ensured, the difference in survey moments in the different countries implied that using the data for straightforward cross-country comparison of agencies was only possible for countries in which the survey years were sufficiently close to each other (for example, the continental countries). Data has hence been mostly used to find cross-country patterns in terms of potential antecedents or effects of autonomy and control, with country dummies reflecting both differences in politico-administrative regime and time (see for a recent overview, Verhoest in this volume and Maggetti and Verhoest 2014).

Although building partially upon the COBRA network, the COCOPS project ("Coordinating for Cohesion in the Public Sector of the Future") differed from the COBRA initiative in that it is a strongly centrally coordinated and funded effort to conduct the same survey in the same time span in a large number of European countries. The objective of the COCOPS Top Public Executive Survey was to capture the views of senior civil servants about recent administrative reforms and their effects across Europe. The survey also collected information on topics such as management practices, the impact of the fiscal crisis, autonomy and politicization etc. It was organized as part of a large-scale European research project (2011–2014), funded by the European Union's Seventh Framework Programme, which looked into the future of public administration in Europe, and the effect of past NPM-Style reforms more specifically (Hammerschmid et al. 2016).

Following input from the 11 research teams involved in the COCOPS project, and an initial screening of existing literature and questionnaires, a central survey design team consisting of representatives from five universities (Hertie School of Governance, Erasmus University Rotterdam, CNRS, University of Bergen, and Cardiff University), first drafted a conceptual model for the survey. This model was subsequently used to make decisions about inclusion and exclusion of topics. The questionnaire was subsequently designed using existing items where possible. The development process required four meetings, after which all research partners were invited to comment. Translation happened by the different national teams, whereby special emphasis was given to the

conceptual equivalence of terminology and national sensitivities. For this process, translation guidelines have been developed, and a translation record was kept to discuss and record tarnation difficulties and subsequent decisions. A test version of the questionnaire was sent to approximately 10 public sector practitioners in each country. The survey was administered centrally by the Hertie School of Governance using Unipark software. Where this was deemed necessary, a paper-based round of data collection has also been organized. National-level datasets were subsequently integrated using a data harmonization protocol

The target population of the survey was the entire population of top public sector executives in central government. This means the entire population received an invitation to fill the survey and no sampling was used. In the case of Germany and Spain, also regional-level executives have been included. A top public executive has been defined as someone occupying the first three hierarchical levels in ministries and agencies and occupying a senior management position. In order to define the exact target population for each country, all national teams collected information about the structure of their public sector. This was done using a standardized excel sheet with a number of questions about both the structure of the public sector and the number of potential respondents. This information was subsequently used to make decisions about inclusion or exclusion of specific individuals or groups.

Initially intended to only cover the ten countries involved in the project, the number of countries rapidly expanded after several universities and government bodies joined the effort. The dataset now contains data on top public executives in 20 European countries: Austria, Croatia, Denmark, Estonia, Finland, France, Germany, Hungary, Iceland, Italy, Ireland, Lithuania, the Netherlands, Norway, Poland, Portugal, Serbia, Spain, Sweden, and the UK. Response rates, and thus data quality, differs across countries. Data for Belgium have been removed from the central database because of insufficient response rates. The final central government dataset contains 7247 observations. A version of the dataset has been made open access and can be downloaded through the GESIS Data Archive for the Social Sciences. The data allows comparisons of data covering all countries (Hammerschmid et al. 2016) as well as within subgroups (Greve et al. 2016).

While the COBRA and COCOPS initiative had a descriptive and explanatory focus, the third initiative has measurement development as its focus. This has a substantial in impact on the methods used (in terms of less demands), but still, there are important issues to be addressed in terms of research design and governance. In 2009–2010, Sangmook Kim and Wouter Vandenabeele initiated a research effort which entailed the development of a measure of public service motivation that could be used cross-culturally and internationally (Kim and Vandenabeele 2010). After all, since public service motivation was to a large extent context-dependent, this affected existing measures of public service motivation which were all aimed at capturing the particularities of the environment as good as possible, causing it to be less effective in doing comparative research. Therefore, the aim was to have a measure that actually measured the same things in all countries involved (measurement invariance, see below).

In total, 12 countries were rallied—South Korea, Belgium, The Netherlands, The United States, The United Kingdom, Denmark, France, Italy, Lithuania, Australia, China, and Switzerland—and 16 scholars were included in this effort. Starting from the broad concept of public service motivation, developed by a series of feedback rounds by all scholars involved. This initial discussion was aimed at developing the concept in a way that would be applicable to all countries involved, as well as at item-specific elements such as values or wording that would be culturally laden and would, therefore, cause responses to differ not in substance but in score (Kim et al. 2013). The survey was conducted in a group of samples of local government employees in all countries (n = 2868).These were chosen because unlike other levels of government, differences in task and context were likely to be minimal—the aim was to survey 'town hall officials' or the administrative positions in local government. Based upon considerations of statistical power, 250 respondents were aimed for in each sample. As can be derived from the total N, this number that was not reached in each country—although some countries overshot, in which case that particular subsample was resampled to obtain a roughly equal number of respondents for each group and avoid over-representation.

The revised dimensions and items were then tested cross-nationally to isolate a set of universal dimensions and determine the extent to which the meaning of these dimensions were shared across the 12 countries studied resulting in a 16 item measure of public service motivation. Although, the results were mixed, in the sense that only for certain comparisons sufficient measurement invariance could be claimed. Still, the final measure could be seen as the best one available to do, for example, smaller scale comparative research.

In general, governance of this research was 'light' in the sense that there was a limited purpose of this study, name to develop the single measure. This enabled us to have a transparent trajectory, with clearly identifiable steps in which clear roles were assigned. Much of the follow-up research based on the dataset was left up to individual or groups of researchers as property rights beyond the single paper fell back to those who had collected the data in the respective countries.

Clearly, the three initiatives are different in their origins and objective, enabling us to learn from these experiences. In the next sections, we turn to some important points of attention in terms of methodological issues and governance approach when setting-up large-scale survey-based comparative research.

59.3 Methodological Challenges in Large-Scale International Survey-Based Research

59.3.1 Doing the International Survey in the Right Way: The Total Survey Error Framework

When doing survey research across countries, one would like to have data from all countries collected at the same time, whereby an identical

methodology with unified standards, procedures, and uniform high quality of survey documentation is used (Kolczynska 2014). This is however rarely the case. Indeed, the recent growth in a number of academic cross-national surveys within public administration has, as discussed by Kolczynska (2014) and Lee et al. (2012), not been accompanied by an improvement in survey quality (see also Smith et al. 2011). This threatens the reliability of comparative research within Public Administration since it introduces several sources of error and bias, thereby reducing data quality. Poor or unknown data quality will in turn raise doubts as to the validity of results, thereby reducing the impact of research, whether on scientific progress or policy recommendations (Kolczynska 2014).

A good way to examine whether survey research is done appropriately is the use of the total survey error (TSE) framework, which assesses "all possible sources of error that can bias survey findings, including sampling error, coverage error, non-response error, measurement error, and processing error" (Lee et al. 2012, 88). This framework thus allows to control for the variety of errors that surround survey research. In order to increase the quality of survey research, one should minimize the total survey error.

The TSE framework (visually presented in Fig. 59.1) includes at the measurement side: validity, measurement, and processing error and at the representation side: sampling error, coverage error, non-response error, and adjustment errors (Groves et al. 2004). Validity is the extent to which measures reflect an underlying concept. By measurement error, we mean the departure from the true value of the measurement as applied to a sample unit and the value provided (Groves et al. 2004). This type of error can have several sources, including question wording. Surveys which are poorly designed that are ambiguous or overly complicated make it difficult for respondents to comprehend and answer adequately (Holbrook et al. 2006). Processing errors refer to errors that can be introduced after the data are collected but before the estimation process (Groves et al. 2004). Sampling error can lead to biased results when the entire target population is not selected and decreases as the sample size increases. A coverage error can exist when a sample frame does not fully represent the total population sampled; leading to a selection bias and consequently generalizability issues (Lee et al. 2012). Non-response error refers to systematic differences in responses between respondents and total sampled persons. Finally, adjustment errors refer to post survey adjustments. These are introduced to reduce coverage, sampling, and non-response errors, but can also increase those (Groves et al. 2004). In what we elaborate below we mainly focus on the issue of response error and the issue of measurement error in the context of large-scale survey-based international comparative research.

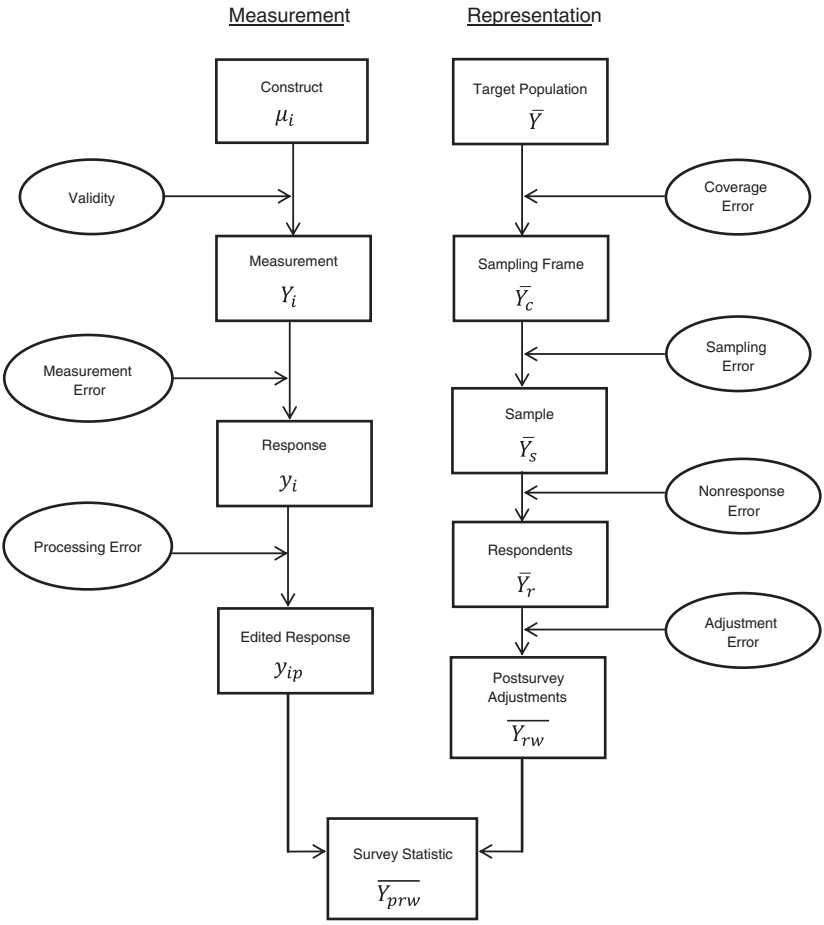

Fig. 59.1 Survey lifecycle from a quality perspective. *Source* Groves et al. (2004)

59.3.2 Are We Surveying the Same Respondents: Non-response Error

Discussing all elements of the TSE framework in detail would lead us too far. Moreover, this has already been thoroughly done for public administration research by Lee et al. (2012). Nonetheless and given our focus on comparative research, an important concern on the representation side of the TSE framework is the possible existence of non-response error. Non-response can be either missing completely at random (MCAR) or being systematic. If non-response is MCAR, then there is no underlying reason why certain sampling units failed to complete the survey. This would relieve the data from

being biased. Systematic bias, on the other hand, occurs when there is some underlying reason why sampling units do not participate in the survey. This biases any results based upon the data to the extent to which respondents differ from non-respondents on variables of importance to the analysis. Moreover, non-response can be divided in unit non-response and item non-response (Loosveldt 2008). Unit non-response refers to the refusal of a person to cooperate with the survey request entirely while item non-response denotes the failure to respond to specific questions in the survey.

Non-response has two main consequences. It reduces the sample size and thus decreases the precision of the estimates and second it affects the sampling design. Current best practices argue that researchers should attempt to maximize response rates and to minimize risk of non-response errors (Olson 2006). However, research (Merkle and Edelman 2002) has called the traditional view into question by showing no strong relationship between non-response rates and non-response bias (Groves 2006). The bias consequently does not only depend on the response rate, but also on the extent to which respondents on non-respondents differ. A higher response rate only minimizes the maximal impact that non-response has under the worst-case scenario. One should therefore not excessively concentrate on high response rates since there is no clear-cut relationship between response rate and non-response bias (Groves 2006; Beullens and Loosveldt 2012). It all depends on how much and on what variables respondents and non-respondents differ.

Dealing with non-response error is already challenging in a single country situation, yet a cross-country setup makes this issue immensely more complicated. Comparing survey results across countries would ideally need minimal non-response error in every country. Since this is not possible, a similar or comparable non-response bias in each country should exist. Given the fact that response rates differ significantly across countries and variables, this is however problematic. Moreover, it is often very difficult to assess non-response bias as it requires either population information with respect to the core variables of a survey, or similar information about the non-respondents. Unfortunately, both are seldom available.

A rather 'easy' way to obtain a view on non-response bias is the comparison between the obtained sample distributions with the population distributions. The source of the expected distributions can be for instance reliable official statistics. A chi-square test can be utilized to test these differences. Yet this technique also has a downside since official statistics have to be available. Representativeness can be claimed based on this variable, however, since information is lacking for other variables the representativeness toward these other variables cannot be tested.

Another approach can exist in response rate comparison across subgroups (Groves 2006, 654–656). One would assume that there is no bias if the hypothesis of no difference between subgroups is not rejected. Following Billiet et al. (2011) one thus asserts that constant response rates across groups

imply non-response bias.[4] Another downside is the fact that non-response bias can be based on variables which are not included in the data.

Alternatively, one could also examine the variation within the existing survey (Groves 2006, 655; Billiet et al. 2009). More in particular estimates from early and late cooperation in a web survey in which several recalls are organized can be compared with each other. It is assumed that late respondents are informative for the final non-respondents (Billiet et al. 2011; Curtin et al. 2000). Future datasets based on a survey with several recalls should include information thereon allowing to calculate this measure.

A better approach, however, exists in the calculation of a Representativity Indicator (also called R-indicator) (see also Schouten et al. 2012). Such an indicator is based on the standard deviation of estimated response probabilities and is defined by (Schouten et al. 2009):

$$M(\rho) = 1 - 2S(\rho)$$

In which $S(\rho)$ is the standard error of the response probabilities. A dataset is said to be representative if all response probabilities are equal. In such a case the standard deviation is equal to zero, and $M(\rho)$ assumes the value 1. Following Schouten et al. (2009) the data is not representative if there is much variation in response probabilities. If the standard error equals 0.5 (maximum value), the value of the R-indicator is equal to zero. Response probabilities have to be estimated using a logit or probit model and should be based on a set of auxiliary variables.

It is important to note that the effect of non-response bias is, however, small when conducting explanatory regression analysis (Billiet et al. 2009; Stoop et al. 2010). However, this does not relieve survey researchers from taking the eventuality of non-response bias seriously since one cannot conclude a priori that the findings are not affected. Future survey research should consequently more frequently report on the existence of non-response bias and, when necessary (e.g., in the case of variation of response rates across important subgroups or departures from distributions on key variables that are known from outside sources for the population), use adjustment methods such as weighting (e.g., the U.S. Federal Employee Viewpoint Survey; see also Loosveldt and Sonck 2008), extrapolation, modeling, and imputation (e.g., Wynen and Oomsels 2013). This should, however, be done with care in order to avoid adjustment errors.

To summarize. Although survey research is being used widely in the field of public administration, information on the quality of the data is often poor or even missing. As discussed by Lee et al. (2012) the TSE framework is a useful tool for categorizing the types of survey error and can help researchers in detecting and reporting possible issues. Hence, future survey research should use the TSE framework as a roadmap and should try to minimize it. Especially, non-response can pose a serious threat to the quality of the data within comparative research. Using tools such as the R-indicator will offer

more transparency into the reliability of survey data, making it easier to judge the quality of the data and the underlying analytical results.

59.3.3 Are We Measuring the Same Things in a Similar Way Across Countries? The Paramount Issue of Measurement Invariance

Another issue in comparative research is the topic of measurement invariance. This is further addressed in the paragraph below. First the general idea is discussed, followed by a typology and a state-of-the-art in public administration.

59.3.3.1 The Comparison of Scores

Most comparative research aims at either describing or explaining variables or phenomena in terms of individual cases. In doing so, invariably the questions on validity are raised along the lines of the following: "are we comparing apples and oranges?". This not only goes for the dependent variable that is used but equally important, it also goes for the independent variable. If one assumes that X influences Y, then it should be established first that X is X in all cases, and the same goes for Y. In quantitative research, this matter of similarities and differences when measuring a feature with distinct groups of respondents is captured very well in theories of measurement invariance.

Measurement invariance addresses the question whether what is measured in one environment is the same as something that is measured in another environment. Despite its apparently straightforward nature, this topic is more complex than it would seem at first glance. There are many characteristics in which a single measure could prove to be invariant and consequently, many types of invariance exist. These can stem from either bias in the construct, bias in the method or bias in items (Jilke et al. 2015). Typologies of measurement invariance vary in their extent (Gregorich 2006; Byrne 1998; Vandenberg and Lance 2000), but they always consist of a set of increasingly restrictive requirements of invariance for specific parameters in a confirmatory factor analysis model. Such a model always assumes that an individual score on a particular item is a function of the latent factor—the broader conception one has about the topic at hand, for example, a feeling of trust—combined with other systematic and non-systematic influence. Is a difference in intercept or starting value, is the influence of the aforementioned latent factor similar over all cases, is the remaining error term similar in size and nature…? These questions all aid in answering the 'apples vs. oranges' question and therefore need to be addressed before any actual comparison takes place.

59.3.3.2 Types of Measurement Invariance

Below, several, increasingly demanding, types of invariance are listed. These also have certain consequences in terms of the subsequent (possible) type

of comparison that can be made. A first type of invariance, which is relatively unimportant in empirical comparative research, is dimensional invariance (Gregorich 2006; Byrne 1998). This type of invariance studies asks the question whether two (or more) measures are invariant with respect to their dimensional make-up. In essence, it investigates whether a multidimensional concept renders the same dimensions in both samples. It is of little use when doing empirical comparative research, but it is a first step in assessing further, stricter types of invariance (Vandenberg 2002).

A second commonly acknowledged type of measurement invariance is configural invariance (Vandenberg and Lance 2000; Schmitt and Kuljanin 2008). The increased restriction here is that the latent factors or sub-dimensions should have the same items across the multiple samples. Evidently, dimensional invariance is prerequisite before assessing this type of invariance. When having assessed this type of invariance, it means that similar instruments can be applied to different cases, as every item loads on the same factor (and none else). Therefore, one can make sample-specific claims, but it does not mean that the samples can be analyzed jointly and meaningfully compared. It could still be the case that the relationship between the latent factor and the item score is different between cases and that a difference between two scores is due to another factor than the substantial difference.

A third type of invariance is metric invariance. This type of invariance assumes that, apart from the previous restrictions, the loadings or λ- values are equal across samples. It assumes equal metrics or scales, which means that different scores on items can be assessed meaningfully across samples (Steenkamp and Baumgartner 1998). This type of measurement invariance is necessary when for example multiple samples are pooled and regression analysis is performed on (a set of) items and be interpreted in the same way. Because this is about correlation, it looks at relative impact—if X changes, then Y also systematically changes; and this is similar in all cases—and it therefore enables to make at least some comparison. Usually, this is done by means of dummy variables for the cases as independent variables in a regression with a metric invariant dependent. A significant dummy would indicate that, opposed to the referent category, a particular dummy (and therefore that particular case) has a positive or negative influence on the level of the dependent variable. The answer to the comparative question would, in this case, be the dummy—and therefore the case; e.g., a particular country—positively or negatively influences the score on the dependent variable. Also, this kind of invariance is required when combining multiple samples of different groups (nationalities but also gender or members of different subsectors) that may have an influence on the actual scores. One has to ask oneself the question "could it be that this group responds differently to my question than other groups"? If the answer is yes, then metric invariance between these groups is required before any further analysis can be undertaken.

A fourth type of invariance is scalar invariance, which not just assumes equal factor loadings, but also equal intercepts. This type of invariance is required when means are compared, as difference in item scores can be due to both the loading and the intercept. This refers to the comparison as we mostly know it: does group A responds differently to a question than group B? This is a rather strict requirement, but nevertheless necessary, since it is the only possible way we can be certain that we are comparing apples to apples: does a score on an item actually means the same in group A as it does in group B? Only when this is answered affirmatively, the actual comparison can be performed.

To assess levels of measurement invariance, the most common method used is comparing measurement models by means of multigroup confirmatory factor analysis (Byrne 1998), although exploratory factor analysis of item-response theory is also sometimes used (Van de Vijver and Leung 2011). For this, regular statistical packages that accommodate confirmatory factor analysis such as LISREL, MPlus, AMOS, STATA or R exist.

59.3.3.3 Measurement Invariance Studies in Public Administration

Although measurement invariance is important when comparing context-sensitive measures, it has been largely neglected in public administration research. This is particularly unfortunate since in the field of public administration, and by extension the fields of political and social sciences, comparative analysis is a very important method. Only recently, more attention has been brought to the issue. Jilke et al. (2015) demonstrated how the presence of measurement variance can cause biased results in two instructive comparisons in the field of public administration.

It is not a surprise that this research is situated mostly within Europe—or at least outside the United States. After all, the need for comparative research is less pressing in a large—at least seemingly—monolithic setting than it is in a patchwork of small countries combined in complex databases. Then again, in the latter environment, the benefit of comparative research is substantially larger and the added value of assessing measurement invariance is therefore much bigger in terms of theory development. This benefit further increases when addressing concepts that theoretically depend on institutions as an antecedent.

Nevertheless, there is a huge lack of interest in the field. To our knowledge, apart from the aforementioned article by Jilke et al. (2015), most research is found in the domain of public service motivation. Here, measurement invariance has been assessed when comparing public service motivation measures in various countries (Kim et al. 2013) or different language groups within one country (Giauque et al. 2011). The former paper is probably a best-developed exercise in assessing measurement invariance so far, as it concerned 12 countries. Compared to other comparative research, the investment in this type of research is relatively low. Each subsample only counted 250 respondents, and only limited attention was devoted to ascertaining representativeness as the main objective was to obtain a sample where

measurement invariance could be claimed—not representativeness. Despite these low costs, the topic remains understudied and deserves more attention.

59.4 Overcoming Limitations of Cross-Sectional International Comparative Surveys

Just as any survey-based research, comparative research based on a single source is prone to be affected by common method bias, causing spurious correlations and therefore questionable findings and conclusions (Podsakoff et al. 2003; Kelman 2015). In statistical terms, this is only one form of endogeneity, but it is one very likely to be present in this type of research (Meier and O'Toole 2013). The easy way to avoid common method bias is to collect data from multiple sources (like using two distinct sets of surveys with different types of respondents, see Favero and Bullock 2015). However, this becomes exponentially more complex in a (international) comparative design. After all, it is one thing to collect data with various national samples as a cross-sectional survey, but it is much harder to find matching datasets, for example, performance of agencies collected by the national government to link to that first comparative dataset, as the outcome set has to match in terms of variables, measures but also timing.[5] As a result, you might end up with a good comparative set for one set of variables, but not being able to match it with the data set on the other variables you are interested in. Since most statistical remedies are deemed not to be effective in coping with common method bias (Favero and Bullock 2015; Jakobsen and Jensen 2015),[6] it is important to consider this carefully at the design stage (see for recommendations, MacKenzie and Podsakoff 2012).

Another strong suggestion for future research is the development of cross-country panel data. Currently, the vast majority of research in public administration is based on cross-sectional data. Although cross-sectional data can be appropriate for studies that examine concrete constructs, employ a diverse array of measurement factors and scales, and are either descriptive in nature of strongly rooted in theory (Rindfleisch et al. 2007), the development of cross-country panel data can have multiple advantages.

A major advantage of such data is the possibility to uncover dynamic relationships. In our opinion, many of the public administration phenomena and their interrelations are inherently dynamic so that most econometrically interesting relationships are explicitly or implicitly dynamic. Consequently, it is not unimaginable that the relationships discussed and examined in our research are subject to causality issues. Again caution is therefore necessary when interpreting results.

Another benefit of panel data exists in the fact that they allow to control for the impact of omitted variables and the issue of endogeneity. "It is frequently argued that the real reason one finds (or does not find) certain effects is due to ignoring effects of certain variables in one's model specification

which are correlated with the included explanatory variables" (Hsiao 2007). Panel data can help solve this issue since information is available on both the intertemporal dynamics and the individuality of the entities and may allow one to control the effects of missing or unobserved variables. A third important benefit of panel data is the fact that they usually contain more degrees of freedom and more sample variability than cross-sectional data, hence improving the efficiency of econometric estimates (Hsiao 2007). Finally, panel data can be extremely interesting for public administration since it allows to accurately evaluating the effectiveness of a certain intervention or policy program. These kinds of data allow to observe the before- and after-effects as well as the possibility of isolating the effects of treatment from other factors affecting the outcome (Hsiao 2007).

59.5 Governance Issues

Organizing large-scale international comparative surveys and governing the collaboration between researchers involves a number of considerations that will greatly improve data quality as well as subsequent analysis. Based on our experience organizing such comparative data collection, the following issues need to be considered.

First, there is the selection of countries. Researchers can decide to cover an entire geographical area, such as the EU, or all South-East Asian countries. When such comprehensive coverage is absent, there need to be good theoretical reasons to include or exclude countries. Otherwise, country selection will be severely criticized by reviewers of articles based on the data: what is the theoretical relevance of writing a paper combining data from, e.g., Spain, the Czech Republic, and Japan? More countries are not always better and make data collection and harmonization difficult. Yet, leaving out large countries from the data collection, such as the UK from European data collection, or the US from a dataset containing mainly OECD countries is also likely to encounter criticism, even when theoretically unfounded.

Second, the best public administration departments are not necessarily the best data collection partners. Thinking in terms of sampling frames and understanding total survey error is a special skill. It therefore makes sense to only involve partners who have had prior survey experience. Also, despite the attractiveness of involving many research partners, be aware that more partners mean more discussions, and almost always longer questionnaires. Be clear about who is in charge of designing the research, and who will be mainly involved in the project for data collection and subsequent analysis. Long questionnaires based on mutual compromises tend not to generate good data. At the same time, researchers should be aware that hot topics in one country may be seen as dull or as too sensitive in other countries. International comparative research should be ambitious, but straying too far from the common denominator comes with risks.

Third, do not assume all public organizations work in the same way as they do in your country. This means that data collection sometimes will have to be adapted to national peculiarities. Think about being granted access, using the hierarchy to promote the survey, using electric vs. paper-based questionnaires etc. Special attention should also go to conceptual equivalence: words, and especially public administration concepts may mean very different things to respondents in different organizations: A concept such as E-government may mean the development of a website in one country, and the development of block chain technology in another.

Fourth, you need either money or a well-integrated intellectual community to pull off a good comparative project, and having both is great. Collecting data takes a lot of time and money, which is not always available to scholars. If you can offer them money or personnel through an international grant, then things become easier. If your team has to apply for money in each individual country, then the project will move very slowly. For smaller data collections, however, having longstanding collaborations and a shared research interest can generate a lot of goodwill, and may help bringing scholars on board, even without money.

Fifth, make good agreements about data use. Who will be allowed to use the data, and who will be allowed to use it first? All datasets have juicy bits that everyone wants to use. Whose names will be on publications using comparative data? Prior to collecting the data, think about a code of conduct outlining the rules of engagement, including data use and data sharing. Sharing data ought to be the norm, but at the same time, those who invest in data collection need to be able to harvest first. A moratorium on data sharing of one or several years makes perfect sense, as does establishing first-use rights for some scholars in the network, either for the entire dataset or for some variables. Data use is an issue, though, that tends to be approached from a fear of losing out or of being scooped. Our experience, however, is that the number of scholars who have time, interest, and skills to exploit international datasets is limited, and that therefore the risk of underutilization of data is probably higher than that of cut-throat competition. Finally, think about the future. Document the research process and make sure all documents and datasets are deposited in a data repository. Too often, data gets lost after key researchers move or retire.

59.6 Discussion: Issues for the PA Community

In this chapter, we discussed a number of methodological and governance challenges when setting-up and managing large-scale international comparative research through the replication of surveys. When successful, such research efforts are very rewarding in terms of scientific insights, practitioners' recommendations, and network building. But, when one wants to structure the data gathering and analysis optimally, both in terms of methodology

and governance, such research is also demanding in terms of resources, time, coordination, and methodological skills. When conducting comparative surveys in international perspective, the design, administration, and exploitation of the survey have to be organized and documented very carefully, taking into account the recommendations listed in this chapter, in order to maximize the scientific quality and to minimize potential skepticism of peers and reviewers. When sufficient care is taken in data gathering and analysis, surveys certainly have their value for descriptive-comparative and measurement-building/testing research, but also for explanatory research. However, in particular, if one wants to complement cross-sectional survey data with other sources of data, or upscale to cross-country panel data, then the need for financial and time resources may become exceedingly high and opportunities may become limited.

Therefore, in our view, European PA scholars—because they are both close to the reality of multiple institutional environments and benefitting the most from this type of research—should take the lead in developing networks that address these questions. In building these networks, it should be explicit what the aim of a particular research effort is and what type of questions will be answered. This focus will enable any comparative effort to maintain the methodological rigor and the governance needed to obtain meaningful results instead reaching sub-optimal outcomes for the time and resources invested.

We conclude this chapter by providing some reflections which might help in such a discussion. First, concerning the kind of knowledge we are seeking for, most of us would probably agree that describing and comparing phenomena is fine, but in order to accumulate knowledge and build theories, we need to understand and explain why these phenomena occur, look the way they do and have certain effects. Long-term data has better cards to help in this endeavor compared to snapshot data, meaning that investments in longitudinal data gathering in the form of panel data should be on the agenda, at least for phenomena which are already well researched (like public service motivation, organizational autonomy).

However, we, being a community of researchers, editors, and reviewers need to keep some realism in our expectations. Still, a lot of research, including such large-scale survey-based international comparative research, is funded by governments and international organizations looking for policy recommendations for pressing problems in a short time span. Matching maximum academic value and practical relevance in that context is not always easy.

Therefore, considering the challenges of large-scale international research, we should consider very carefully which topics and research questions really need such an approach. In most of the cases, we do not really need such large-scale international comparisons. We should not do such research just because it is fancy and great-looking, fun to do, or great for our reputation. If we want to know the impact of administrative traditions and politico-administrative regimes on public administration phenomena or measurements,

we should first check whether the following two strategies are sufficient, just because they are far more easy to organize, to control for methodological quality, and to add other research methods and data besides survey. One strategy is simply to run a focused, conceptually well-developed, well-managed single country survey, which is later on as part of a pure replication study done in another country. The other strategy is to set-up a small international network (3–4 partners) with well-chosen countries, representing relevant independent variables, carefully selected partners. In such studies, surveys are preferably combined with other methods like document analysis, case studies, and experiments in order to build a case for causal relationships.

In the end, either of such studies could be the stepping stone for topics and research questions that really need large-scale survey-based comparative research. Given that this will only be relevant for a limited set of topics and research questions, we should think about our strategy toward funders and governments. As optimal international large-scale research needs much financial and time resources, we should lobby for more international funding by addressing EU, co-funding by multiple national research councils or the OECD or the World Bank as venues. Moreover, considering the resource intensiveness of international surveys at such a scale, there is the tendency to increasingly use secondary data stemming from, e.g., government surveys, like staff/citizen satisfaction surveys. Using such data has its advantages, but there might be problems in terms of data quality (TSE—see above), a poor match between research concepts and the gathered data, as well as an incomparability between such surveys from different countries. However, as a PA community, we should develop a strategy to plea with the appropriate platforms (like EUPAN or the European Commission) for conceptual and methodological optimization as well as international coordination and equalization of such national government-based surveys, combined with open access to such data. In the long run, that is probably only the feasible and enduring strategy to get long-term panel data on some core public administration issues in international perspective.

Notes

1. This chapter is partially based on a presentation held at the Study Group Chairs Meeting at the EGPA conference in 2015 by Gerhard Hammerschmidt, Wouter Vandenabeele and Koen Verhoest.
2. Information about the network and survey can be found on http://www.publicmanagement-cobra.org/.
3. The involved European countries are Belgium, Norway, Ireland, Italy, The Netherlands, Germany, Austria, Switzerland, Portugal, France, Sweden, Denmark, Finland, Lithuania, and Romania. Later the survey was also held in Spain and in the German Länder.
4. This relies on the missing at random (MAR) assumption.

5. Moreover, governments are not particularly well-known for their ability to match their research and data-collection efforts to that of other governments; it could very well be that question are differently framed, or surveyed at another point in time.
6. One potentially promising method is to extract a single method factor that is used as a control (see Favero et al. 2016). This is however a very data-intensive method and it puts a severe test on data in that it is likely to throw away part of a real relationship. So it is biased to generate null results.

References

Bach, T., Ruffing, E., & Yesilkagit, K. (2015). The differential empowering effects of Europeanization on the autonomy of national agencies. *Governance: An International Journal of Policy, Administration and Institutions, 28,* 285–304.

Beullens, K., & Loosveldt, G. (2012). Should high response rates really be a primary objective? *Survey Practice, 5*(3), 1–5.

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahwah: Lawrence Erlbaum.

Billiet, J., Davidov, E., & Schmidt, P. (2011). *Cross-cultural analysis. Methods and applications.* New York: Routledge.

Billiet, J., Matsuo, H., Beullens, K., & Verhovar, V. (2009). Non-response bias in cross national surveys: Designs for detection and adjustment in the ESS. *Research and Methods, ASK, 18,* 3–44.

Curtin, R., Presser, S., & Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. *Public Opinion Quarterly, 64,* 413–428.

Favero, N., Meier, K. J., & O'Toole, Jr, L. (2016). Goals, trust, participation, and feedback: Linking internal management with performance outcomes. *Journal of Public Administration Research and Theory, 26*(2), 327–343.

Favero, N., & Bullock, J. B. (2015). How (not) to solve the problem: An evaluation of scholarly responses to common source bias. *Journal of Public Administration Research and Theory, 25,* 285–308.

Giauque, D., Ritz, A., Varone, F., Anderfuhren-Biget, S., & Waldner, C. (2011). Putting public service motivation into context: A balance between universalism and particularism. *International Review of Administrative Sciences, 77,* 227–253.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44,* 78.

Greve, C., Lægreid, P., & Rykkja, L. H. (2016). *Nordic administrative reforms. lessons for public management.* Basingstoke: Palgrave.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* Hoboken, NJ: Wiley.

Groves, R. M. (2006). Non-response rates and non-response bias in household surveys. *Public Opinion Quarterly, 70,* 646–675.

Hammerschmid, G., Van de Walle, S., Andrews, R., & Bezes, P. (Eds.). (2016). *Public administration reforms in Europe. The view from the top.* Cheltenham: Edward Elgar.

Holbrook, A. L., Cho, Y. I., & Johnson, T. P. (2006). *Extreme response style: Style or substance.* Paper presented at the annual meeting of the American Association for Public Opinion Research, Montreal, Canada.

Hsiao, C. (2007). Panel data analysis—Advantages and challenges. *Test, 16*(1), 1–22.
Jakobsen, M., & Jensen, R. (2015). Common method bias in public management studies. *International Public Management Journal, 18,* 3–30.
Jilke, S., Meuleman, B., & Van de Walle, S. (2015). We need to compare, but how? Measurement equivalence in comparative public administration. *Public Administration Review, 75*(1), 36–48.
Kelman, S. (2015). Letter from the editor. *International Public Management Journal, 18,* 1–2.
Kim, S., & Vandenabeele, W. (2010). A strategy for building public service motivation research internationally. *Public Administration Review, 70*, 701–709.
Kim, S., Vandenabeele, W., Wright, B. E., Andersen, L. B., Cerase, F. P., Christensen, R. K., 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,* 79–102.
Kołczyńska, M. (2014). Representation of Southeast European countries in international survey projects: Assessing data quality. *Research and Methods, 23*, 57–78.
Lee, G., Benoit-Bryan, J., & Johnson, T. P. (2012). Survey research in public administration: Assessing mainstream journals with a total survey error framework. *Public Administration Review, 72,* 87–97.
Loosveldt, G. (2008). Nonresponse error. In P. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 533–536). Thousand Oaks: Sage.
Loosveldt, G., & Sonck, N. (2008). An evaluation of the weighting procedures for an online access panel survey. *Survey Research Methods, 2*(2), 93–105.
Maggetti, M., & Verhoest, K. (2014). Unexplored aspects of bureaucratic autonomy: A state of the field and ways forward. *International Review of Administrative Sciences, 80*, 239–256.
MacKenzie, S. B., & Podsakoff, P. M. (2012). Common method bias in marketing: Causes, mechanisms, and procedural remedies. *Journal of Retailing, 88,* 542–555.
Meier, K. J., O'Toole Jr, L. J. (2013). I think (I am doing well), therefore I am: Assessing the validity of administrators' self-assessments of performance. *International Public Management Journal, 16*, 1–27.
Merkle, D., & Edelman, M. (2002). Nonresponse in exit polls: A comprehensive analysis. In M. Robert, D. A. Groves, J. L. Eltinge, & J. A. Roderick Little (Eds.), *Survey nonresponse* (pp. 243–257). New York: Wiley.
Olson, K. M. (2006). Survey participation, nonresponse bias, measurement error bias, and total bias. *Public Opinion Quarterly, 70*, 737–758.
Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.
Rindfleisch, A., Malter, A. J., Ganesan, S., & Moorman, C. (2007). Cross-sectional versus longitudinal survey research: Concepts, findings, and guidelines. *Journal of Marketing Research, 45*(June), 261–279.
Schouten, B., Cobben, F., & Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology, 35,* 101–113.
Schouten, B., Bethlehem, J., Beullens, K., Kleven, O., Loosveldt, G., Luiten, A., et al. (2012). Evaluating, comparing, monitoring and improving representativeness of survey response through R-indicators and partial R-indicators. *International Statistical Review, 80,* 382–399.

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

Smith, S. N., Fischer, D. S., & Heath, A. (2011). Opportunities and challenges in the expansion of cross-national survey research. *International Journal of Social Research Methodology, 14,* 485–502.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25,* 78–90.

Stoop, I., Billiet, J., Koch, A., & Fitzgerald, R. (2010). *Improving survey response. Lessons learned from the European social survey.* Chichester: Wiley.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5,* 139–158.

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedure. In D. Matsumoto, & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17–45). New York: Cambridge University Press.

Verhoest, K., Peters, G. B., Bouckaert, G., & Vermeulen, B. (2004). The study of organizational autonomy: A conceptual overview. *Public Administration and Development, 24,* 101–118.

Verhoest, K., van Thiel, S., Bouckaert, G., & Lægreid, P. (Eds.). (2012). *Government agencies in Europe and Beyond: Practices and lessons from 30 countries.* Basingstoke: Palgrave Macmillan.

Wynen, J., & Oomsels, P. (2013, September 11–13). *Analyzing inter-organizational trust: How to obtain trustworthy results?* Edinburgh: EGPA.

Yesilkagit, A. K., & van Thiel, S. (2012). Autonomous agencies and perceptions of stakeholder influence in parliamentary democracies. *Journal of Public Administration Research and Theory, 22,* 101–119.