# Towards good enough measurement : sick leave statistics as a case of the measurement challenges in comparative public performance

# *Towards good enough measurement – sick leave statistics as a case of the measurement challenges in comparative public performance*

*Abstract*

*This paper puts forward suggestions how to deal with typical problems found in international comparisons of governance, illustrated by the case of sick leave statistics. We analysed sick leave data and meta-data of employees in the public administration of six European countries and regions. The purpose of the analysis is to learn about the challenges of, and potential solutions for international comparisons of performance. The study demonstrates that problems of reliability and validity can be reduced, provided that sufficient data and meta-data are available. These results suggest that although "perfect comparisons" are difficult to achieve, work can be done to make measurement "good enough".*

Keywords: comparative public administration, rankings, performance measurement, secondary data analysis

## Introduction

> *"Railing against the rankings will not make them go away; competition, the need to benchmark, and indeed the inevitable logic of globalization make them a lasting part of the academic [and political; added by the authors] landscape of the 21st century. The challenge is to understand the nuances and the uses — and misuses — of the rankings."*
>
> (Altbach, 2010)

Rankings are everywhere: both of governance as a whole and of specific policy sectors. For higher education alone there exist at least three world-wide rankings: the Academic Ranking of World Universities (ARWU, the "Shanghai Rankings"), the QS World University Rankings, and the Times Higher Education World University Rankings (THE).

Based on these rankings, many students choose their university and study programmes. In a similar way, international governance rankings and comparisons play a role in legitimizing decision-making (e.g. Gormley and Weimer 1999, Almeida and 15 others 2001, Andrews 2008, Arndt 2008, Hood 2007, Pollitt, Bouckaert, and Van Dooren 2009, Pollitt 2011, Erkkilä 2015). The reactions to the results of the Programme for International Student Assessment (PISA) study 2011 in Germany, and consequent efforts to reform the education system represent just one, illustrative example (Raidt 2009). But also within the field of comparative public administration there has been a boom in international rankings. Examples are the World Bank "Worldwide Governance Indicators" (WGI), the competitiveness indicators of the IMD Business School, the World Economic Forum (WEF) rankings, and the OECD's Government at a Glance study (OECD 2011, 2013). Thus, both individual decisions and public policies rest on international rankings. We therefore should not treat their quality and contents lightly (Van de Walle, Sterck, Van Dooren, and Bouckaert 2004).

It has been demonstrated that in fact many of these indicator sets face considerable problems of validity and reliability (Almeida and 15 others 2001, Van de Walle, Sterck, Van Dooren, and Bouckaert 2004, Arndt and Oman 2006, Kaufmann, Kraay, and Mastruzzi 2006, van de Walle 2006, Hood, Dixon, and Beeston 2008, Luts, Van Dooren, and Bouckaert 2008, Kaufmann, Kraay, and Mastruzzi 2009, Van Dooren 2009, Andrews, Hay, and Myers 2010, Langbein and Knack 2010). Luts et al. (2008) and Hood et al. (2008) analysed existing government and governance rankings in terms of their conceptualisation, methodology, sources, and indicators used, as well as the resulting validity and comparability of these rankings. Whereas Hood et al. (2008) included rankings which measure both policy performance (for example education, health) and government performance (public administration), Luts et al. (2008) focused on

governance rankings only. All of the rankings analysed by these scholars have major deficiencies, especially with regard to the conceptualisation and actual measurement. The underlying concepts are either not defined at all, or only very vaguely (for example Worldwide Governance Indicators of the World Bank Institute). The choice of indicators also remains mostly unexplained, and in many cases the rankings are based purely on perceptual data (for example Worldwide Governance Indicators). Furthermore, sample sizes are very often too small to be called representative. For instance, the International Institute for Management Development relies for the soft data of its World Competitiveness Yearbook on the responses of only 44 people to draw conclusions about Belgium (Luts et al. 2008: 106–107). Nevertheless, it seems reasonable to assume that these rankings are "here to stay". Consequently, we have to choose between disagreeing with, and neglecting them, or "going ahead with *good enough* measurement" (Van de Walle 2009: 53, emphasis in the original).

Good enough measurement might be a solution for the dilemma between "academically correct measurement" and "measurement as a tool for improvement" (Van de Walle 2009: 53). Van de Walle (2009: 43) contends that especially in the public sector we have to accept "certain imperfect data as good enough for measurement". We have to base any effort of comparing public administration performance internationally on certain basic agreements, such as, what is public administration, what is the measured subject, while at the same time we have to be aware that these definitions are context- and time-dependent (Pollitt 2008). Thus, on the one hand we want meaningful indicators that reflect the complexity of what is measured. On the other hand, indicators ought to be simple enough to be useful – both for policy-makers and accountability (Van de Walle 2009: 48). So how do we get to such a "good enough measurement"?

This question is the one the present paper wants to answer. Based on differences found in international rankings, we put forward suggestions for reducing the problems of international (in)comparability of indicators. We illustrate our suggestions with examples from a comparison of sick leave statistics of public entities of a number of Western European countries and regions. Sick leave is an indicator, which is used in governance rankings to assess working conditions in governments (for example OECD 2011), the latter indicating an aspect of performance within public administration. By choosing sick leave as an example, we also build on Van Dooren, De Caluwe, and Lonti (2012), who suggest measuring public administration performance by looking into performance indicators of, for example, budgeting, human resource management (HRM), and open government practices.

The case of sick leave indicators can be considered representative for other indicators in the sense that there exist no binding definitions of it, nor standards of how to measure it. Consequently, the findings and suggestions presented in this article can be applied to a variety of other (public administration) indicators, where similar conditions hold.

After sketching an analytical framework for our analysis, we present the results of our study. In our conclusions we reflect on what the case of sick leave statistics can tell us about international comparative measurement in general.

## 1. Analytical framework

Typical measurement problems have been analysed by various scholars, and based on their findings we will construct our framework. Table 1 below lists the main challenges of international rankings and comparisons found in the academic literature. As can be seen, these challenges centre around issues of reliability and validity. For example, Hood et al. (2008) examined 14 international governance rankings against six criteria of

validity and reliability. Luts et al. (2008) found nine typical problems with governance rankings, all related to reliability and validity issues. Finally, Van de Walle (2006) assessed three governance rankings against conceptual validity and data quality. For our purposes, we selected two validity criteria against which the sick leave indicators will be analysed: Concept validity, which describes the consistency of underlying methodologies and definitions; and measurement validity, which is linked to measurement and data transparency. The data reliability (that is disaggregation of data, in our cases) will also be addressed. We consider these to be among the most pressing issues, as is also reflected in the scholarly attention dedicated to them.

*** Insert table 1 about here. ***

These criteria are related in particular to Luts et al.'s (2008) and Hood et al.'s (2008) findings: Among the nine typical problems Luts et al. found with governance rankings, the problem of overaggregation corresponds to Hood et al.'s criterion of availability of disaggregated data. Similarly, the problem of conceptual deficiencies is the equivalent of Hood et al.'s validity criterion of a coherent underlying methodology and theory. Measurement and data transparency, finally, is reflected in two problems found by Luts et al. (2008): A lack of transparency, and measurement errors. The problem of (in)comparability of results in governance rankings is the overarching challenge of all attempts at international comparisons, and as such, also plays a vital role in our assessment of international comparative performance measurement.

Sick leave is a prominent indicator when measuring public administration performance: The OECD includes it in their "Government at a Glance" study (OECD 2011), and the Common Assessment Framework (CAF) of the European Institute of Public Administration suggests it as one of their "people results" (CAF 2006). Also in academia,

absenteeism is an important topic in relation to an organisation's performance (for example (Osterkamp and Röhn 2007, Cristofoli, Turrini, and Valotti 2011, Bierla, Huver, and Richard 2013, Gosselin, Lemyre, and Corneil 2013). Hence, we consider sick leave as a typical performance indicator.

More importantly in our context, however, is sick leave a relevant indicator for addressing the measurement problems of public administration performance. We will thus use sick leave statistics of a few European countries and regions to illustrate the challenges and possible improvements of public administration performance measurement. As such, assessing and improving the reliability and validity of sick leave statistics permits us to draw lessons for international comparative measurement in general.

## 2. Analysis of sick leave statistics as a measurement case

We compared the sick leave statistics of the core public administration of six small, Western European countries and regions, to situate our study also in a multi-level context: Austria, Bavaria (DE), Denmark, Finland, Flanders (BE), and the Netherlands. The choice of these countries and regions is based on their geographical proximity, similar administrative traditions, established statistical systems, socio-demographic homogeneity, as well as on data accessibility. However, in our discussion we are not going to go into detail about the respective cultures and potential explanations for differences in sick leave quotas, as this is not the main purpose of the paper. With this selection of Northwest-European administrations, we hope to minimise variation on structure and culture, in order to discuss variation in the measurement challenges properly.

Data and information on the methodology of collecting the data were available online on the sites of the governments or statistical offices in almost all cases. Austria and Bavaria publish a detailed report online. Data for Finland were received by e-mail, as the data are usually only available for employees of State agencies and departments. The Flemish government provides rather limited information online, but the missing information and data have been provided by getting in e-mail contact with employees working in the respective departments.

Due to the lack of international comparative datasets, and the consequent lack of international standards for measuring sick leave, we were confronted with several issues we had to overcome in order to make the data obtained (relatively) comparable, and as such limit the overarching problem of international comparability.

**Concept validity: Consistent underlying methodologies and definitions**

With respect to consistent underlying methodologies and definitions, we found two main issues which reflect the problem of the lack of, or differences in, underlying methodology and theory.

*Definitions of sick leave and public administration*

To begin with, we looked at the basic concepts of sick leave and public administration. While the definition of "sick leave" was comparable across all countries and regions examined, this was not the case with public administration.

*Sick leave* is defined as *absence from work due to sickness, which does not include pregnancy or maternity leave.*[1] Hence, regarding the definition of sick leave, we find the validity criterion of a consistent underlying concept to be present.

An interesting side-note is the case of Denmark: In its statistics Denmark distinguishes between "own sickness" and "sickness of own children". This distinction is not explicitly made in any of the other countries and regions examined, and we could not find any comments regarding the issue in their sick leave reports. It might be interesting to look into national and regional regulations with regard to absence due to the sickness of employees' own children, when assessing a country's sick leave quota.

More challenging was the definition of *public administration:* We are specifically interested in the statistics of "core national and regional Public Administration", which covers the administration of core departments and ministries, but excludes defence, justice, and public services such as schools, universities, health care, as well as municipalities.[2] This "narrow" concept of public administration however, was not much used in the countries examined:

Generally, in all regions and countries "public administration" is used to refer to the broad notion of public administration, with the notable exception of Flanders and Finland. Nevertheless, Austria includes separate data for the core public administration in its sick leave report, and the National Statistics Bureau of Denmark has data for core public administration for the year 2010. For Bavaria and the Netherlands data for core

[1] In Bavaria, the sick leave definition explicitly excludes stays at health resorts, which appears to be a very regional definition.

[2] This definition follows the definition of "general public services" according to the Classification of the Functions of Government (COFOG). (*http://unstats.un.org/unsd/cr/registry/regcst.asp?Cl=4*, accessed 12th November 2013).

public administration could not be calculated, and as such, we had to leave them out of any comparative analysis.

*Population and calculation*

Moving to the concrete data gathering for sick leave statistics, we found differences regarding the population included in the calculations, and the days on which the calculation of the sick leave quota is based. These issues are also closely linked to the second validity criterion of transparency. We will turn to this in the following section, and look more closely into matters concerning the underlying methodology and concepts now.

The first important difference among cases is the question of *who is included in the calculation*: While in some of the countries and regions looked into calculations are based on fulltime equivalents, in others the data of all employees actively working at the time of data collection were included, regardless whether they worked full- or part-time. This has quite a substantial impact on the results: Three employees, one working fulltime, the other two part-time, equal three workers in headcounts. Yet in fulltime equivalents (fte) they count as two.

A second difference concerns *calculations based on calendar or working days*. Are sick leave calculations based on calendar days or actual working days? Are average (paid) holidays included or excluded? Regarding the first question, Statistics Netherlands (Centraal Bureau voor de Statistiek) has demonstrated the implications of using calendar or working days in its "standard for sick leave registration" (Centraal Bureau voor de Statistiek (CBS) 2005).

If only fulltime employees are taken into account, and only sick leave of longer duration (six weeks, in their study) considered, then there is no difference in the sick leave quota whether calendar or working days are used for the calculation:

*Calculation based on calendar days:*

*(6 weeks * 7 days) / (1 * 365 days) * 100% = 11.5%*

*Calculation based on working days:*

*(6 weeks * 5 days) / (1 * 261 days) * 100% = 11.5%*

Nonetheless, a difference does appear if the duration of sick leave cannot be counted in whole weeks (CBS 2007: 25).

*Calculation based on calendar days:*

*(6 weeks * 7 days + 4 days) / (1 * 365 days) * 100% = 12.60%*

*Calculation based on working days:*

*(6 weeks * 5 days + 4 days) / (1 * 261 days) * 100% = 13.0%*

Hence, the result depends on the number of weekend days included in the days absent due to sickness above and beyond "full" weeks: If there are more Saturdays or Sundays in these additional days of sick leave, then the quota based on calendar days will be higher. If, on the other hand, the additional days taken off are working days, then the method based on working days will lead to a higher quota.

If we further consider sick leave at an amalgamated level, that is sick leave based on the whole population (of those working in public administration), then the difference between calendar and working days disappears. Counting days of sickness, the result based on calendar days will be 7/5 times higher than the result based on working days. When calculating the percentage, both methods will yield the same result, as the

numerator in the calendar days method will be 7/5 times higher than that of the working days method. The opposite is true for the denominator (CBS 2005: 26–27).

Yet the difference is still relevant in regard to part-time workers: Part-time workers have to be included in the calculations in such a way that their sick leave is not unintentionally weighted higher or lower. As such, the aim is to include them proportionately in the calculations (pro rata). When applying the working days method, both these conditions are fulfilled, while when applying the calendar days method, part-time workers are treated as if they were fulltime workers. If only the quota and the days absent due to sickness are considered, no difference will be noted. Nevertheless, the average duration of sick leave will seem to be higher among part-time workers, as less cases of sick leave will be registered (CBS 2005: 27-28).

Consequently, the working days method leads to a more reliable result: Although there is a difference in actual sick days between the two groups (fulltime and part-time workers), this difference is proportional to the so-called part-time factor[3]. With regard to the quota, there is no difference between fulltime and part-time workers (CBS 2005: 28).

Despite this being so, the calendar days method is still in frequent use. To overcome the problem of part-time workers being given undue weight in the calculation, the CBS (2005: 28–29) has suggested that both methods be combined by calculating sick leave as a proportion of actual calendar days missed, based on the part-time factor (whereby the part-time factor of individual cases has to be applied to individual sick days, and not an average part-time factor to the total number of sick days of all part-time workers).

[3] The part-time factor is calculated as *work hours per week / work hours full time employment* (CBS 2005, p. 3).

*Practical relevance*

In our cases, we found evidence of both individual methods: Austria calculates sick leave by referring to calendar days. Bavaria, Flanders, and Finland base their calculations on working days, which are defined as *calendar days minus weekends and public holidays.* For the Netherlands we found a remark that departments are free to choose which method they want to apply. However, as the CBS (2005) states, if based on calendar days, they use the alternative method which combines both approaches, so that the end result does not differ. Denmark bases its calculation on working days, which are defined as calendar days minus *weekends, public holidays, and annual (paid) leave*. This does not remain without consequences for the sick leave quota: It would actually be lower than that presented. A direct comparison with countries which calculate their sick leave data without taking into account personal holidays must accordingly be carried out with caution.

Another case for having to adjust the calculation of the sick leave quota represents Austria. It calculates its sick leave quota based on

*calendar days: total days of sick leave / (number of employees * 365) * 100*.

Yet this leads to a slightly lower percentage than the more usual method of calculation (days of sick leave per employee / number of working days per employee), which for 2010, would have resulted in a quota of 5.5 per cent instead of the given 5.0 per cent:

*Total days of sick leave / (number of employees * 365) * 100*

*902 475,3 / (49 241,90 * 365) * 100 = 5.02%*

*Sick leave per employee / working days per employee * 100*

*13.85 / 252 * 100 = 5.50%*

The actual number of days absent due to sickness are however calculated in working days.

In conclusion, we do not find a consistent underlying concept of public administration across the countries and regions we investigated. We furthermore found differences regarding the population included in the calculation, and the calculation of the sickness quota itself. Accordingly, comparing data on public administration internationally has to be done with caution, and only after a thorough investigation of the specific underlying concepts and methodologies.

**Measurement validity: Measurement and data transparency**

The measurement validity criterion, measurement and data transparency, is closely linked to the previous one. In addition to the differences encountered and outlined in the previous section, we were also confronted with issues concerning data and measurement transparency in the cases examined. The underlying concepts and the methodology on which the calculations are based, are very often not stated clearly. Consequently, in particular the difference between the calculation of sick leave based on fulltime equivalents or headcounts of employees is very difficult, if not impossible, to correct for. This is especially the case if we do not have access to the raw data.

The availability of disaggregated data, in fact the overaggregation of data, is an issue we had already come across at the very beginning of our data collection process. In none of the regions and countries examined could we get access to the disaggregated data. This however leads to comparability problems, as has been shown in the previous sections.

To illustrate our point, let us go back to the difference between calculations based on fulltime equivalents or headcount. If we wanted to correct for the differences resulting

from these two measurement methods, we would need to know exactly how many people are employed fulltime and part-time in public administration, ideally with an explanation (at individual level) as to whether part-time employees work every weekday for less than a full day, or whether they work only some days of the week, but for the full length of each day (which, according to the CBS (2005: 27-28), also has an impact on the results). This information is not available. As a result, we can only point out that differences in the population included in the calculation lead to differences in the results.

But more generally, as has become obvious in the previous sections, the reliability and validity of international comparisons not based on a uniform definition of the concepts measured, stand and fall with the availability of disaggregated data. It is not necessary, nor is it feasible, to introduce new, uniform definitions. What is, however, necessary and feasible to achieve "good enough measurement", is to provide access to the data and meta-data used. Hence, validity and reliability of comparative sick leave statistics is impaired unless we can shed light on the potential differences of the cases examined.

**An additional challenge: Completing missing data**

Apart from the criteria already discussed, we were confronted with an additional challenge, which is the one of how to deal with missing data. Trend data are a very useful source for performance measurement, as they allow us to trace the performance development of the indicators examined. Therefore, when aiming at assessing the performance in the field of sick leave, having sick leave data for more than one year at our disposal would give us a clearer picture of the matter at hand.

However, for Denmark we were only able to obtain the sick leave quota of the core public administration for 2010, whereas for earlier years we have the quota of the whole

public administration. Accordingly, we were able to complete the missing data: The data for both the core and the whole public administration are publicly accessible via Statistics Denmark. The respective tables, FRA05 (Absence by sector, sex, cause of absence, age and indicator of absence) for the broad public administration, and FRA10 (Absence in governmental sector by region, sex, cause of absence, industry (DB07) and indicator of absence) for the core public administration, are based on the same population, and as such, can be combined.

Thus, we assume that the proportion between the sickness quota of the core public administration and that of the total public administration is stable across time. This assumption enables us to calculate the sickness quota for the core PA:

*Sickness rate core / sickness rate total = 3.75 / 3.55 = 1.06*

*Sickness rate core = 1.06 * sickness rate total*

*2009: 1.06 * 3.57 = 3.68*

*2008: 1.06 * 3.52 = 3.73*

*2007: 1.06 * 3.87 = 4.10*

Nonetheless, it has to be borne in mind that these numbers are in fact assumptions. We cannot say for sure that these numbers reflect reality. They do however show relative developments, and as such can be used for illustrative purposes. Furthermore, this example of a recalculation of existing data shows a way to tackle comparability problems based on differing concepts and definitions. However, this was possible only because there was sufficient necessary information about the data and meta-data available. In summary, this example illustrates the importance of the availability of data, and the interlinkage between the validity and reliability criteria applied in this study.

## 3. Results: Comparability across countries is limited

Taking account of all the measurement differences outlined above, we arrive at three out of six originally included countries and regions which can be compared with each other with some confidence: Flanders, Denmark, and Finland. All three regions base their sick leave measurement on fulltime equivalents, and use the same method to calculate their sick leave quota (see Table 2). Nevertheless, as Denmark calculates the quota based on working days minus holidays, the actual sick leave quota is even lower than presented in our calculation. Due to limited data availability, it was not possible to re-calculate the data so that this difference would become visible.

*** Insert Table 2 about here ***

If the difference between headcount and fulltime equivalent were ignored, we would also be able to compare these three regions with Austria, where equivalent data for core public administration were available for two years (2008 and 2010). However, in the same way as the quota for 2010 was re-calculated to make it comparable with calculations undertaken in other countries, resulting in a slightly higher value, the quota for 2008 would have to be adjusted. As we were not able to obtain the actual days defined as sick days in 2008, this step could not be taken, and it needs to be borne in mind that the stated quota for 2008 should actually be somewhat higher.

## 4. Discussion of results: Diverging results in comparison with the data of the OECD

The presentation of the actual sick leave data, rather than the methodology, leads to a remarkable finding: Our data differ noticeable from the ones presented by the OECD, (2011). The OECD's Government at a Glance study of 2011 contains a section on "Working conditions in central government" (OECD 2011: 132–133), which includes sick

leave as one of the indicators. Nevertheless, we find comparability across countries to be limited as there is no standardised definition of "government", nor "public administration", which does influence the results as shown in the figure below. We found the difference between the results of the data collected by us, and the data presented by the OECD in their "Government at a glance" study (OECD 2011) very intriguing. While the figures for Finland and Denmark do not differ very much, the difference regarding the Austrian and Dutch data is notable. We found sick leave in Austria to be 15 per cent higher, and in the Netherlands 80 per cent (see figure). Flanders does not appear in the figure, as we do not have the corresponding data of the OECD. Important to note is, however, that the data in the figure below are not comparable across countries, due to significant differences in the measurement methodology. Rather, our results, which we obtained by consulting the data of the respective national statistic offices, can be compared with the results of the OECD for each country. We have not conducted an in-depth analysis of the reasons for these differences, and consequently limit ourselves to suggesting a few assumptions which might be worth considering.

*** Insert Figure about here. ****

As we have outlined above, the biggest challenge in comparing sick leave quotas across countries is the variety of definitions and measurements surrounding this issue. In particular the definition of public administration differs widely. Some countries take only direct administrative staff into consideration, others hold a broader definition and also include employees not directly employed in administration, or do not differentiate

between administration and executive. We have shown that these differences in definition have a considerable impact on the results.

However, as there is no standard definition of public administration and/or (central) government OECD countries adhere to, countries provide their data to the OECD according to their individual definitions. For example, when the OECD asks for "civil servant" data for sick leave, it will receive these data based on the definitions of the respective countries: If a country's definition of civil servants includes employees in the "executive", this country will provide sick leave data accordingly. Other countries, which apply narrower definitions of civil service or central government, that is restricted to public administration (administrative staff in ministries and departments), will equally provide their corresponding data. Consequently, the results of the OECD are based on a variety of definitions and measurements. We do not know which country applies which definition and which measurement. As such, a direct comparison across countries is limited. As we have demonstrated above, however, this issue could be partially solved by providing more information about the data and meta-data used to compile these rankings.

So why do we find a visible difference between our data and that of the OECD in the case of the Netherlands and Austria, but not in the case of Denmark and Finland? The answer might be the following: Denmark and Finland base their measurement on the same definition of "core government" (that is for us, public administration) as we have used for our results. In contrast, the "standard definition" of the Netherlands and Austria of core government comprises a wider group of employees, which leads to the mentioned variations of results when correcting for these differences.

Therefore, the added value of our study lies in having been able to go deeper into the data collection procedure in a small number of countries. In doing so, we are able to correct differences (where adequate data are available), and try to avoid comparing apples and oranges. We are by no means claiming to have achieved perfect comparison. We see it rather as a modest attempt to raise awareness about the difficulties one has to face when comparing these kind of data across countries without having a standard definition which every country under investigation adheres to. At the same time we wanted to come up with suggestions to reduce these problems, especially in view of the multitude of rankings available and used.

**Conclusion**

This paper shows that there are substantial differences regarding a variety of concepts and definitions concerning sick leave statistics, which make an international comparison challenging. The most frequently found problems concern differences in, or lack of, transparency of underlying concepts and definitions, which limit comparability across cases.

The fact that there are no uniform, standard definitions of the main concepts, such as public administration, add to this difficulty. However, finding a common standard is everything but an easy task. Van de Walle (2009: 52) even claims that "we cannot measure government because we cannot define it." Or, put differently, as we cannot define government, we cannot measure it. He brings the discussion down to a distinction between an econometric approach to the issue at hand, and a public administration approach. While public administration might have to be satisfied with "good enough measurement" which is at times based on what Van de Walle calls arbitrary decisions, econometricians should pay more attention to underlying conceptualisations and

theoretical models, especially when dealing with such complex concepts as public administration and governance.

Thus, if relevant institutions, such as the OECD or Eurostat, engaged in finding generally accepted definitions and principles for registering sick leave statistics, this could, on the one hand, contribute significantly to the comparability of these data across countries. On the other hand, this would have major implications for the national statistics gathering, as national systems most likely would have to be changed substantially. As the countries we have compared are relatively homogeneous, the differences in their measurement methodology are most likely not rooted in different measurement cultures, but rather in their institutional path dependency. Thus, this type of changes would imply the investment of not only financial, but also human and time resources. Nevertheless, this could be a valuable endeavour for the long term. In the short term, however, we have to deal with what we have.

Yet we have shown that most of the mentioned problems can be reduced provided we have access to the necessary data and meta-data. This does not only apply to sick leave statistics, but to statistics and rankings of all kinds, as we have shown elsewhere (Hoffmann & Van Dooren, 2013). Regarding data availability, we detected different policies in the countries and regions under examination. While some provide detailed data and analyses on their websites, others maintain a rather restrictive policy, and provide data only upon request or with very limited additional information, such as population included in analysis, calculation method etc. In line with this, we also want to stress the importance of transparency within government and towards citizens. Data availability and transparency go hand in hand in the sense that transparency is a precondition for the usefulness of available data: If data are presented in a way difficult

(for laymen) to understand, this might impact negatively on the willingness to access them – government will remain a mystery for most of its citizens.

In fact, when reviewing the data collection and analysis process, we can conclude that it all links back to transparency: The more information and the more data are available (and comprehensible), the more measurement problems can be reduced.

As we have sketched briefly in the introduction, the lack of reliability and validity of governance rankings is alarming. Nevertheless, there are ways to overcome these limitations. This paper has aimed at illustrating one of these ways, and as such, contributes to the literature on rating and ranking assessments with empirical illustrations.

Furthermore, the results suggest two main implications for comparative public administration: On the one hand, there is a need for more caution and a more critical look when analysing available data. When comparing actual results of international comparisons of performance, public administration, governance (see, for example, Afonso, Schuknecht, and Tanzi 2005, Jonker 2012, Freistein 2015) researchers and practitioners have to be aware of the various differences across cases, beginning with the underlying concepts and definitions, and including the methodology applied, as well as inherent cultural or structural aspects. Hence, regarding practitioners, we call for more transparency and openness of the data included in the studies. As concerns researchers, we encourage them to a close look at these data, especially when comparing across countries, regions, or sectors.

On the other hand, these results also suggest that quantitative comparisons of public administration are feasible. If we want to keep on using them however, we should find a middle ground on basis of which we can come to comparable results. In other words, it

is time to get closer to a "good enough measurement" (Andrews 2008, Van de Walle 2009, Andrews et al. 2010). In this line, there is also reason to call for more multi-method approaches (Perry 2012, Pollitt 2013): By combining qualitative and quantitative methods, the data can be put in their respective contexts, and a more nuanced picture of the issues at stake can be drawn.

## Bibliography


Afonso, A., Schuknecht, L., & Tanzi, V. (2005). Public sector efficiency: An international comparison. *Public Choice*, *123*(3), 321–347. http://doi.org/10.1007/s11127-005-7165-2

Almeida, C., & 15 others. (2001). Methodological concerns and recommendations on policy consequences of the World Health Report 2000. *The Lancet*, (357), 1692–1697.

Altbach, P. G. (2010, November 11). The state of the rankings. Retrieved from http://www.insidehighered.com/views/2010/11/11/altbach

Andrews, M. (2008). The good governance agenda: Beyond indicators without theory. *Oxford Development Studies*, *36*(4), 379–407. http://doi.org/10.1080/13600810802455120

Andrews, M., Hay, R., & Myers, J. (2010). Can governance indicators make sense? Towards a new approach to sector-specific measures of governance. *Oxford Development Studies*, *38*(4), 391–410. http://doi.org/10.1080/13600818.2010.524696

Arndt, C. (2008). The politics of governance ratings. *International Public Management Journal*, *11*(3), 275–297. http://doi.org/10.1080/10967490802301278

Arndt, C., & Oman, C. (2006). *Uses and abuses of governance indicators*. Paris: Development Centre of the Organisation for Economic Co-operation and Development.

Bierla, I., Huver, B., & Richard, S. (2013). New evidence on absenteeism and presenteeism. *The International Journal of Human Resource Management*, *24*(7), 1536–1550. http://doi.org/10.1080/09585192.2012.722120

Centraal Bureau voor de Statistiek. (2005). *Berekening van verzuim. NVS standaard voor verzuimregistratie: Nationale verzuimstatistiek*. Den Haag: Centraal Bureau voor de Statistiek (CBS). Retrieved from http://www.cbs.nl/NR/rdonlyres/412FAADD-209E-4511-8B1A-EFF4A8A89AB2/0/2007verzuimstandaard.pdf.

Cristofoli, D., Turrini, A., & Valotti, G. (2011). Coming back soon: Assessing the determinants of absenteeism in the public sector. *Journal of Comparative Policy Analysis: Research and Practice*, *13*(1), 75–89. http://doi.org/10.1080/13876988.2011.538542

Erkkilä, T. (2015). Global governance indices as policy instruments: Actionability, transparency and comparative policy analysis. *Journal of Comparative Policy Analysis: Research and Practice*, 1–21. http://doi.org/10.1080/13876988.2015.1023052

Freistein, K. (2015). Effects of indicator use: A comparison of poverty measuring instruments at the world bank. *Journal of Comparative Policy Analysis: Research and Practice*, *0*(0), 1–16. http://doi.org/10.1080/13876988.2015.1023053

Gormley, W. T., & Weimer, D. L. (1999). *Organizational report cards*. Cambridge, Mass: Harvard University Press.

Gosselin, E., Lemyre, L., & Corneil, W. (2013). Presenteeism and absenteeism: Differentiated understanding of related phenomena. *Journal of Occupational Health Psychology*, *18*(1), 75–86. http://doi.org/10.1037/a0030932

Hoffmann, C., & Van Dooren, W. (2013). *Regional benchmarking of public administration performance. Towards a construction of an international comparable dataset.* Antwerpen: SBOV.

Hood, C. (2007). Public service management by numbers: Why does it vary? Where has it come from? What are the gaps and the puzzles? *Public Money & Management*, *27*(2), 95–102. http://doi.org/10.1111/j.1467-9302.2007.00564.x

Hood, C., Dixon, R., & Beeston, C. (2008). Rating the rankings: Assessing international rankings of public service performance. *International Public Management Journal*, *11*(3), 298–328. http://doi.org/10.1080/10967490802301286

Jonker, J.-J. (2012). *Countries compared on public performance : A study of public sector performance in 28 countries*. The Hague: The Netherlands Institute for Social Research/SCP.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2006). *Governance Matters V: Governance Indicators for 1996–2005*. Washington, D.C: World Bank.

Kaufmann, D., Kraay, A., & Mastruzzi, M. (2009). *Governance Matters VIII: Aggregate and individual governance indicators, 1996-2008*. Retrieved from http://papers.ssrn.com/abstract=1424591

Langbein, L., & Knack, S. (2010). The Worldwide Governance Indicators: Six, one, or none? *Journal of Development Studies*, *46*(2), 350–370. http://doi.org/10.1080/00220380902952399

Luts, M., Van Dooren, W., & Bouckaert, G. (2008). Internationale rangschikkingen gerangschikt. Retrieved 5 June 2012, from https://lirias.kuleuven.be/handle/123456789/204162

OECD. (2011). *Government at a Glance 2011*. Paris: OECD.

OECD. (2013). *Government at a Glance 2013*. Paris: OECD.

Osterkamp, R., & Röhn, O. (2007). Being on sick leave: Possible explanations for differences of sick-leave days across countries. *CESifo Economic Studies*, *53*(1), 97–114. http://doi.org/10.1093/cesifo/ifm005

Perry, J. L. (2012). How can we improve our science to generate more usable knowledge for public professionals? *Public Administration Review*, *72*(4), 479–482. http://doi.org/10.1111/j.1540-6210.2012.02607.x

Pollitt, C. (2008). *Time, policy, management: governing with the past*. Oxford ; New York: Oxford University Press.

Pollitt, C. (2011). 'Moderation in all things': International comparisons of governance quality. *Financial Accountability & Management*, *27*(4), 437–457. http://doi.org/10.1111/j.1468-0408.2011.00532.x

Pollitt, C. (Ed.). (2013). *Context in public policy and management: the missing link?* Cheltenham, UK ; Northampton, MA, USA: Edward Elgar.

Pollitt, C., Bouckaert, G., & Van Dooren, W. (Eds.). (2009). *Measuring government activity*. Paris: OECD.

Raidt, T. (2009). *PISA: Katalysator im bildungspolitischen Paradigmenwechsel. Dimensionen des Wertewandels im Bildungswesen*. Heinrich-Heine-Universität, Düsseldorf. Retrieved from http://docserv.uni-duesseldorf.de/servlets/DerivateServlet/Derivate-14323/Bildungsreformen-PISA-Raidt_A1b.pdf

Van de Walle, S. (2006). The state of the world's bureaucracies. *Journal of Comparative Policy Analysis: Research and Practice*, *8*(4), 437–448. http://doi.org/10.1080/13876980600971409

Van de Walle, S. (2009). International comparisons of public sector performance. *Public Management Review*, *11*(1), 39–56. http://doi.org/10.1080/14719030802493254

Van de Walle, S., Sterck, M., Van Dooren, W., & Bouckaert, G. (2004). *What you see is not necessarily what you get. Een verkenning van de mogelijkheden en moeilijkheden van internationale vergelijkingen van publieke sectoren op basis van indicatoren* (No. D/2004/10107/011). Leuven: Instituut voor de Overheid.

Van Dooren, W. (2009). A politico-administrative agenda for progress in social measurement: Reforming the calculation of government's contribution to GDP. *Journal of Comparative Policy Analysis: Research and Practice*, *11*(3), 309–326. http://doi.org/10.1080/13876980903220751

Van Dooren, W., De Caluwe, C., & Lonti, Z. (2012). How to measure public administration performance. *Public Performance & Management Review*, *35*(3), 489–508. http://doi.org/10.2753/PMR1530-9576350306

X: Deleted for blind review.

Tables and figure – Towards good enough measurement – sick leave statistics as a case of the measurement challenges in comparative public performance

Table 1: Overview of challenges of international rankings

| Challenge | Authors |
|---|---|
| **Measurement transparency** | Almeida & 15 others 2001; Arndt & Oman 2006; Hood et al. 2008 |
| **Data transparency** | Almeida & 15 others 2001; Arndt & Oman 2006; Hood et al. 2008; Kaufmann et al. 2006, 2009 |
| **Conceptual deficiencies** | Almeida & 15 others 2001; Langbein & Knack 2010; Luts et al. 2008; Van de Walle et al. 2004; Van Dooren 2009 |
| **Overaggregation** | Almeida & 15 others 2001; Andrews et al 2010; Kaufmann et al. 2006, 2009; Luts et al. 2008; Van de Walle 2006 |
| **Availability of disaggregated data** | Arndt & Oman 2006; Hood et al. 2008; Van de Walle et al. 2004 |
| **Coherent methodology & theory** | Andrews et al. 2010; Arndt & Oman 2006; Hood et al. 2008; Van de Walle et al. 2004 |
| **Lack of transparency** | Almeida & 15 others 2001; Arndt & Oman 2006; Luts et al. 2008 |
| **Measurement errors** | Arndt & Oman 2006; Hood et al. 2008; Kaufmann et al. 2006, 2009; Luts et al. 2008 |
| **Incomparability of results/data** | Arndt & Oman 2006; Van de Walle et al. 2004; Van de Walle 2006; Luts et al. 2008 |
| **Data availability** | Almeida & 15 others 2001 |
| **Representativeness** | Almeida & 15 others 2001; Luts et al. 2008; Van de Walle 2006 |
| **Perceptual vs factual data** | Andrews et al. 2010; Arndt & Oman 2006; Kaufmann et al. 2006, 2009; Luts et al. 2008; Van de Walle 2006; Van de Walle et al. 2004 |
| **Timelineness of data** | Van de Walle et al. 2004 |
| **Selection bias** | Andrews et al. 2010; Arndt & Oman 2006; Luts et al. 2008; Van de Walle 2006; |
| **Stability of measures over time** | Hood et al. 2008 |
| **Stability of units of comparison over time** | Hood et al. 2008 |

Table 1: Overview of sick leave measurement differences across countries

| Country / region | head count | full-time equivalent | core PA | broad PA | calendar days | working days |
|---|---|---|---|---|---|---|
| **Austria** | x | | x | x | x | |
| **Bavaria** | x | | | x | | x |
| **Denmark** | | x | only for 2010 | x | | minus annual (paid) leave |
| **Finland** | | x | x | | | x |
| **Flanders** | pro rata | x | x | | | x |
| **Netherlands** | | x | | not education/health | x | x |

Figure: Sick leave in days per employee for 2010

Sources: OECD (2013); AT: Bundeskanzleramt (2012); FI: database "Tahti" of the Finnish government (made available by e-mail); DK: Statistics Denmark; NL: NEA 2010