

A Method to Deal with Dissimilar Circumstances of Public Organizations in Performance Comparisons: Evidence from Dutch Prisons

Toon Molleman¹ & Peter G.M. van der Heijden^{2,3}

¹ Scientific Research Centre (WODC) of the Ministry of Security and Justice, the Netherlands

² Utrecht University, the Netherlands

³ University of Southampton, United Kingdom

Correspondence: Toon Molleman, WODC, Ministry of Security and Justice, Postbus 20301, 2500 EH Den Haag, the Netherlands. Tel: 316-5287-7102. E-mail: t.molleman@minvenj.nl

Received: July 19, 2013 Accepted: August 2, 2013 Online Published: September 29, 2013

doi:10.5539/par.v2n2p1

URL: <http://dx.doi.org/10.5539/par.v2n2p1>

Abstract

What are the methodological requirements of performance measures? To what extent can managers influence performance scores and do they have similar organizational circumstances? What is needed for sound and fair comparisons between organizations? In this article, a step-by-step plan for performance comparisons between organizations is proposed in which both administrative and methodological challenges are addressed. The plan is illustrated with two performance measures derived from the Dutch prison system. Performance analysts may use the plan to analyze performance of public organizations.

Keywords: performance comparisons, performance score adjustment

1. Introduction

For several decades, companies and public services have used performance measures to obtain information about the achievements of operational management (Radnor & Barnes, 2007). An additional aim for public services is to 'shape and manage incentives for individual and/or organizational behavior, and to promote transparency and accountability to the public of government activities and their outcomes' (Barnow & Heinrich, 2010: p. 62). Performance measures are used to compare achievements of organizations and help decision makers to allocate human and material resources as well as budgets (March & Sutton, 1997; Poister, 2010). Differences in performance scores may stimulate inferior performing organizations to make efforts for better performance, e.g. via benchmarking principles (see Camp, 1989). In his consolidated view of reasons for measuring performance, Pidd (2013) suggests six categories: planning and improvement, monitoring and control, evaluation and comparison, accountability, financial budgeting and planning, and individual performance management.

Administrators usually assess the performance of an organization with the application of monitoring systems. By doing so, an important point of interest is regularly overlooked. That is, it should be considered whether or not performance scores *purely* reflect the efforts of organizational management and its staff. Depending on the performance measures used, the scores may partly be a result of the given circumstances of an organization (Nyhan & Martin, 1999). When performance scores are not fully related to these efforts, it is advisable to make statistical adjustments. If we omit such considerations there is every chance that the wrong benchmark is indicated and organizations exchange 'best' practices that may lead to worse performance.

It is worthwhile to quote Gaes et al. (2004: pp. 51-52) with regard to inmate misconduct in American corrections to illustrate the significance of performance score adjustment: 'Comparing prisons with unadjusted rates assumes that a naïve comparison is warranted. This is naïve because the substantive assumption was that prisons do not differ in ways other than the ability of management to generate incentives to encourage good behavior from inmates, or disincentives that discourage inappropriate behavior. The assumption of prison equality, except for differences in management effectiveness, is most likely not true. Prisons hold different types of inmates, even when they are purportedly inmates of the same security level.'

In other fields, such as the fields of medical care and education, adjustments of performance measures are sometimes applied as well. Typical performance measures that are adjusted are of a logistic nature, like mortality rates in hospitals (e.g., Drösler et al., 2012; Silber, Rosenbaum & Ross, 1995; Landon et al., 1996; Staiger et al., 2009). These statistical adjustments are performed with the use of prior evidence on the relation between the performance measure and certain selected factors. In the field of education, research into student performance found (next to individual student variation) systematic variation between countries, neighborhoods, schools and classes (Fung et al., 2010; Meyer, 1997). Fried et al., (2002) showed that adjusting performance scores in nursing homes changed the rankings of the homes dramatically. We may therefore say that the added value of adjustment has been scientifically established in several professional fields. Although some techniques for the adjustment of performance measures emerged recently, their use is anything but common practice (Barnow & Heinrich, 2010). Moreover, the application of so-called Bayesian statistical techniques in organizational science is scarce (Kruschke, Aguinis & Joo, 2012). As such, in this article we propose a systematic step-by-step plan for performance comparisons that includes such adjustment techniques. In explaining these steps, we start with organizational goals and end up with a comparative performance ranking. In the plan, we make use of elements proposed by the above authors and also introduce some new elements.

The following steps are considered:

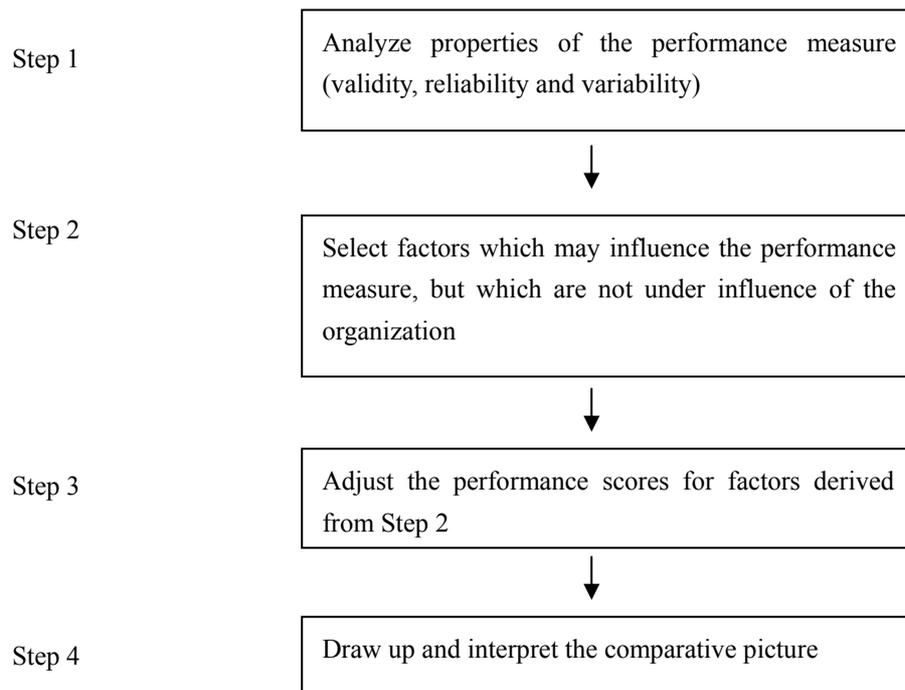


Figure 1. Step-by-step plan for performance comparisons

The steps demand various skills of an analytical, methodological and administrative nature. In the remainder of this contribution, we will work out these four steps by means of two examples of performance measures from the Dutch prison system. For the benefit of readability, in every step we describe the methods as well as the empirical results. Before doing so, we briefly describe two issues: a characterization of the Dutch prison system and the two illustrative performance measures is provided; second, the issue of *complex data* on which performance measures are often based is described.

1.1 Illustrative Example: Dutch Prison System

The Dutch prison system is a publicly run service that operates by using an organizational performance monitor. In this monitor several measures are used to evaluate the goals of the service on an annual basis. Important goals concern safety, humanity and re-integration of inmates. These goals are not unique to The Netherlands. To monitor these goals, prison systems around the world keep a diversity of information up-to-date, e.g., escapes, completed educational courses by inmates and sick leave of staff. This type of information is derived from several sources such as bookkeeping of incidents, staff and inmate surveys and audits. In the Netherlands, correctional facilities

differ in, for example, their architectural design, cell capacity, the use of cell sharing, amount of available staff per inmate and inmate population.

In this contribution we use two empirical examples of performance measures for the annual evaluation of the 45 prisons of the Dutch prison system in 2006-2007 (in the remainder of this article, not all of the 45 prisons are represented due to missing data). Since we have access to multiple data sources concerning *safety*, we use performance measures regarding this specific goal. On the one hand, we present *violent incidents* between inmates that led to punitive segregation. *Violent incidents* are derived from bookkeeping at the prison *unit* level. On the other hand, we discuss the performance measure *staff's feelings of safety*. This measure is collected using a staff survey on the individual level and consists of a scale of five items measuring the extent to which staff feel safe in the institution and think their working conditions guarantee safety. The scale construction is adopted in Appendix I. We realize that performance measurement does not necessarily need to focus on outcomes or outputs; measures of cost effectiveness, efficiency, process and input may be of importance as well (Poister, 2010). However, this is beyond the scope of this article.

1.2 Complex Data

For both examples data are not collected at the prison level, but at a lower level in the organization. Yet it is the prison level on which we want to make comparisons, so there is a hierarchical structure in the data collected on the performance measures. The registrations of *violent incidents* are collected for 125 units in 29 prisons; thus, here the data have two hierarchical levels. Correctional officers (1,689), who work in 172 prison units that are part of 43 prisons, report *Staff's feelings of safety*; thus, here the data have three hierarchical levels. Performance measures are not necessarily collected at the organizational level on which the performance comparisons will be made. This is not unique to prisons; hospitals (with patient evaluations), police forces (with evaluations of citizens) and secondary schools (with results of pupils) aggregate lower level data to the institutional level for the aim of comparisons. It is important to note that characteristics of *individuals* may be related to the performance scores on the *institutional* level. Possibly, certain clients (e.g., inmates, pupils or patients) may be assigned to a specialized institution (e.g., high security prison, elementary school or academic hospital). Individuals within a particular institution tend to behave in a similar way because they share experiences and interact with each other (Raudenbusch & Bryk, 2002). As a result, individual data are not statistically independent observations, an issue that must be taken into account in a statistical analysis. First, the result of dependent observations is that the *effective* sample size is reduced when respondents or units have similar scores within a group (Leyland & Groenewegen, 2003). Due to the similarity in the scores per group, a standard analysis of the data could easily yield significance in too many cases if the hierarchical structure is ignored (due to an overestimated precision). It is important to take the hierarchical structure into account in one overall regression model to circumvent biased estimates and resulting inferences. Second, in regression analyses we might also want to investigate the contribution of factors at different levels (e.g., individual client characteristics and the typology of the institutional building). Multilevel regression modeling allows for this and accounts for the risk of inferential problems due to dependent observations (Gaes et al., 2004; Snijders & Bosker, 2012). We will use the multilevel model in each of the four steps discussed below.

2. Step 1: Performance Measures Require Certain Properties

The choice for specific performance measures depends on the formulated organizational goals. Once goals are established for an organization, there must be attention paid to a sound operationalization into performance measures (unfortunately, this issue falls outside the scope of this article). Once the measures are selected, whether or not the measures are informative has to be investigated. To assess the usefulness of a measure for performance comparisons, their properties must be considered. Useful performance indicators are meaningful and understandable for decision makers, they are balanced and comprehensive in the light of the organizational goals, can be timely provided, and are actionable for decision makers (Poister, 2010). Furthermore, they must be guarded from perverse effects (see for an overview: Smith, 1995). Here we confine ourselves to the methodological criteria of performance measures, namely (1) validity and reliability, and (2) variability.

2.1 Validity and Reliability

Validity refers to the question: Do I measure what I intend to measure? The concept of validity was introduced in 1950's and addresses construct validity or the extent the operationalization represents the phenomenon we want to investigate (Cronbach & Meehl, 1955). The operationalization must pay attention to all relevant elements of the phenomenon (by means of face or content validity tests). Furthermore, with criterion-oriented validity it is tested how an operationalization performs in comparison to some criterion (how well does our operationalization predict

the phenomenon, distinguish groups, match with or diverge from other sources?). Finally, we also distinguish the extent to which research results may be generalized (external validity).

Reliability deals with the problem: If I measure again under identical circumstances, do I find identical performance scores? In social science, reliability is synonymous with consistency. The question to be answered is whether two or more measures give a consistent view of the phenomenon (Cronbach, 1947). Reliability may be tested by taking measures at different moments in time (test-retest reliability), by different observers (inter rater reliability), and by different measures within one testing instrument (internal consistency reliability).

2.2 Illustrations for Reliability and Validity

The validity of the *violent incidents* measure depends on skills and loyalty to reporting rules of staff. We may assume that experienced staff detects incidents more easily and that loyal staff neatly record every event. Some prison staff are simply more effective in identifying deviant activities while others are 'more efficient in obtaining inmate compliance without resorting to written 'tickets' for insubordination' (Gaes et al., 2004: p. 50). Furthermore, coder or deliberate errors should be considered because staff may have an incentive to underreport incidents. Moreover, their 'reporting loyalty' may decline when they know the measure is used for performance monitoring (Poister, 2010; Hood, Dixon & Beeston, 2008). As these skills and the loyalty of staff vary between prisons, the measure may not reflect organizational performance in terms of safety; rather, it reflects skills and behavior of staff. Therefore, for the *violent incidents* measure to be valid for the comparison of performance, it is necessary that the detection and recording skills of staff are comparable over the prisons units (i.e., the level of data collection).

In the organizational performance monitor used in the Dutch prison system, measures are identically defined and prevailing laws and prison rules are identical for all prisons. This promotes the trustworthiness of the performance measure concerning *violent incidents*. In addition, we asked the business controllers of every prison to assess the accuracy of the *violent incidents* measure. Because the measure refers to incidents that led to punitive segregation (and serious events will not easily be overlooked), the controllers were convinced of the faithful representation of the records. In one prison a business controller assessed the records of violent incidents as not reliable. Therefore, we exclude one prison from further analyses and proceed under the assumption that the *violent incidents* measure is valid and reliable.

With respect to the staff survey measure of *staff's feeling of safety*, we have a reasonable survey response of 63% and good representation of the population if we look at several background variables. We tested representation based on the background variables age, sex, working hours, and tenure. The responding officers are not different from the non-responding officers for these variables. Furthermore, validity is enhanced because the anonymity of respondents was guaranteed and survey questions were phrased in such a way as to prevent socially desirable answers. The survey scale (4 items) is reliable with a Cronbach's α of 0.86. The final aim is to compare prisons on the basis of the surveys of the correctional officers. In view of the large sample size for the correctional officers (N=1,689), the parameters estimated at the prison level have a small standard error and are therefore more reliable. Therefore, for *staff's feeling of safety* we also proceed under the assumption that there are no serious problems regarding validity and reliability.

2.3 Variability

Variability refers to whether the organizations differ on the performance measure. When a performance measure does not uncover differences between the organizations, comparisons between those organizations are not meaningful. The mechanism of comparing and contrasting performance implies, among other things, that when differences are found between the organizations, inferior performers get an incentive to improve. Thus, we need performance measures that vary between different organizations (Laird & Louis, 1989); variance of the performance measure can be used here for assessment. However, when measures are not collected at the organizational level, the level at which performance comparisons are made, we must take the hierarchical structure of the data into account.

Multilevel models allow us to estimate variances that can be allocated at each level of the hierarchical data. For this purpose, the models have to be estimated in the absence of explanatory variables (intercept only models). Using these variances the intraclass correlation coefficient (ICC) can be derived. The ICC expresses the degree of resemblance between observations belonging to the same organizational unit (Snijders & Bosker, 2012). Thus, the ICC can be used to assess the amount of variance that exists in the performance measure at the organizational level. An exceptional case is that the observations turn out to be independent; in that case, there is no variance that can be allocated at the organizational level of the data, the ICC is zero and we can conclude that organizations do not differ on this measure.

2.4 Illustrations for Variability

The *violent incidents* measure and the *staff's feeling of safety* measure both have a hierarchical structure that we will take into account in answering the question regarding whether or not there is variability in these measures on the prison level. In the case of *violent incidents*, the dependent variable is a count that we model using a Poisson regression, where the size of the prison unit is taken into account by using $\log(\text{number of inmates})$ as an offset. The *violent incidents* measure is collected on the prison *unit* level. As we want to make evaluations on the prison level, the corresponding multilevel model has two levels. A likelihood-ratio (LR) test comparing a two-level model with a one-level model shows that a two-level Poisson model on the *violent incident* measure gives a significantly better fit than an ordinary Poisson model.

The ICC for *violent incidents* amounts to .097, showing that 9.7% of the total variance can be allocated on the prison level. In Poisson distributions with two-levels, this is calculated by $\sigma^2_{\text{prison}} / [\sigma^2_{\text{prison}} + (\pi^2/3)]$. We conclude that the requirement that the prisons differ on the performance measure is fulfilled. Note that, as only 9.7% of the variance in the performance scores is related to the prison level, the prison management can only be partly held responsible for the variability in the performance measure.

For *staff's feeling of safety*, we apply linear regression procedures. The survey is measured on the *individual* staff level. The model distinguishes three levels; next to the individual level, we consider the prison unit level (level 2) and the prison level (level 3). The likelihood ratio-test shows that a 3-level model is appropriate. For the *staff's feeling of safety* measure, the ICC is .071 on the prison level (and .073 on the prison unit level), thus, 7.1% of the variance can be allocated at the prison level. In linear distributions with 3 levels, this is calculated by $\sigma^2_{\text{prison}} / (\sigma^2_{\text{prison}} + \sigma^2_{\text{unit}} + \sigma^2_{\text{ind}})$. It follows that, even though most of the differences in this performance measure are on the level of the correctional officers, there is a sufficient amount of variance at the prison level to use this measure for a further comparison of prisons.

3. Step 2: Selection of Possible Non-Discretionary Factors

The question addressed in the first step is whether there is variance in performance measures at the organizational level. In other words, do the organizations differ in terms of the performance measure averaged over the lower level(s) in the organizations? In this second step we investigate whether this variance can be attributed to managerial effort. Some 35 years ago Charnes, Cooper and Rhodes (1978) presented methods for 'objectively' determining efficiency in so-called *decision-making units* while taking stock of multiple inputs and outputs. Their first Data Envelopment Analysis model (DEA) assumed that the inputs and outputs were entirely under managerial control. In the last decade, there is growing attention for the problem that a score on a performance measure is not necessarily a result of managerial effort in full (e.g., Fried et al., 2002; Camanho, Portela & Vaz, 2009).

According to Tsai and Bridges (2011) the variability of performance scores between organizations fall into three components: systematic variance, valid variance and random variance.

- *Systematic variance* refers to non-discretionary (ND) factors, namely those factors that are out of the sphere of influence of management. The term is often used in DEA literature for uncontrollable variance (Fried et al., 2002);
- *Valid variance* concerns factors that are within control of management and staff of the organization. This part of the variance represents the differential *performance* of the organizations;
- *Random variance* can be captured with the disturbance term of a stochastic model.

Tsai and Bridges (2011) propose to adjust performance measures for systematic variance, so that only the valid variance remains. Performance measures adjusted in this way can then be used validly to compare organizations.

The question then is, which ND factors (the systematic variance) should be considered to adjust the performance scores? The decisions regarding which factors are non-discretionary and which are not, is a decision that has to be made before statistical analyses are conducted. Ideally, the *selection* of ND factors is well considered and does not follow mere guesswork of the analyst or depend purely on available data. The selection process of ND factors is an opportunity for involving stakeholders (i.e., managers, analysts, and administrators) with the aim to create managerial commitment in the organizations under comparison. When stakeholders agree on the selection made, they will acknowledge the performance score adjustments more easily and have more trust in meaningful comparisons.

3.1 Illustrations for Step 2

For the selection of ND factors, we invited a delegation of prison managers (6 prison directors and the head of service) to become members of an expert group. We prepared this meeting by making a list of potential ND factors

from the literature. Experts could add factors to the list in case they missed relevant ones. Once the list of potential ND factors was complete, the panel had to decide which factors to include. A moderator guided the session and steered towards agreement among the experts. Two questions were put central: 1) Is the factor relevance to safety in the prison? and, 2) As a prison manager, can you influence the factor? The experts graded the former question on a 4-point scale from 'not relevant' to 'very relevant,' and the latter question on a 4-point scale from 'not influenceable' to 'very influenceable.' In all cases, the experts reached consensus and were able to make a joint assessment. We assigned factors as non-discretionary when the expert group assessed them as at least a little relevant and not influenceable. The following factors were labeled as very relevant to safety in a prison *and* were assessed as not influenceable (ND factors) for a prison manager:

- Individual characteristics of inmates (e.g., age, sex, sentence length, and criminal history)
- Staff-inmate ratio
- Cell sharing
- Prison capacity
- Building (architectural design) and;
- Regime

The expert group also reached an agreement on factors that could potentially lead to valid variance, such as the *composition* of inmate characteristics within prison units, human resource factors, and leadership. These *changeable* factors represent the differential performance of the prisons that can be attributed to management and staff. Performance analysis should therefore not adjust for these factors.

Before we go into adjustments of performance scores whether or not there might be a specific systematic variance component, namely systematic measurement bias, must be considered. As stated with respect to reliability in Step 1, the use of records of *violent incidents* might be problematic because these depend on the official who observes and reports (which is not the case with staff surveys). This potential measurement bias should also account for in an adjustment model for *violent incidents* by including the factors 'tenure' and 'loyalty to registration rules' of staff in the particular prison (unit).

Since both performance measures are not derived at the individual *inmate* level, individual characteristics of inmates (i.e., age, sex, sentence length, and criminal history) cannot be included as ND factors in the models presented in Step 3. Therefore, we adopt three aggregate measures of these characteristics since prison management cannot select his or her inmate supply. We aggregate the characteristics to the *prison level* because prison management may influence the composition of inmates on the *unit level* (e.g., by spreading or concentrating specific inmates in certain units within the prison). Unfortunately, a variable for 'loyalty to work instructions' is not available. The other factors mentioned are included in the multilevel regression models presented hereafter. We apply a significance threshold of $\alpha = 0.10$ for prison level variables for inclusion in the models since we have modest statistical power on the highest level.

4. Step 3: Adjustment of Performance Scores

There is no established method to adjust performance scores, but performance score adjustment mostly implies the following: applying regression models, the ND factors are used as explanatory variables to predict the performance measures. Thus, the *predicted* performance measures take into account differences between the organizations in terms of the ND factors. The difference between the *observed* performance measure and the *predicted* performance measure, called *residuals* in a regression model context, reveals the differences that could not be explained by ND factors. Interest goes out to these differences.

However, as we are interested in these differences on the organizational level, we have to study residuals at the organization level. These can be obtained with a multilevel regression model that provides residuals for each level of the hierarchical structure when a so-called random intercept for organizations is included in the model. The current step (Step 3) deals with the prediction of performance measures by the ND factors. Studying the residuals at the organizational level is the subject of Step 4.

4.1 Illustrations for Step 3

We continue with the analysis of the two Dutch prison performance measures. In Step 1 it became clear that there is variation between prisons in the two measures *violent incidents* and *staff's feeling of safety*. For the former measure, a multilevel Poisson regression model is estimated with two hierarchical levels: prison units and prisons. In the latter measure, a multilevel linear regression model is estimated with three different levels in the data: correctional officers, prison units and prisons.

First, the results are provided for the multilevel Poisson regression model fitted for the performance measure *violent incidents*. This model fits a random intercept for the prisons. Since we have 29 prisons and assumed 7 relevant ND factors on that level, statistical power becomes an issue. Therefore, we eliminate prison level variables post hoc that do not reach an alpha level of 0.10. As a consequence, the inmate characteristics at the prison level and the double bunking variable are dropped (not applicable, see Table 1). The regression model in Table 1 shows that most of the ND factors in the model reach statistical significance.

Table 1 shows that the factor *regime* has a significant relation to the amount of *violent incidents*. On the prison level, the building types are related to the prevalence of incidents, as well as staff inmate ratio. However, the interpretation of the connections found is beyond the scope of this contribution. In Step 1 we detected that 9.7% of the variance is attributed to differences between prisons (this is the model without explanatory variables; the percentage represents the unexplained variance at the prison level). With the inclusion of the ND factors in the model above, 5.1% of variance between prisons remains unexplained. Thus, the ND factors explain 4.6% of the variance. We conclude that ND factors are of influence, but there is also sufficient variance (namely 5.1% of the total variance) that represents the effort of the prison (management).

Table 1. Multilevel Poisson regression model for dependent variable *violent incidents*. N=125 units in 29 prisons with *unit capacity* as exposure variable, Wald $\chi^2(11) = 82.13$, Prob = 0.00. NA = not applicable.

Violent incidents	B	S.E.	Sign
<i>Level 1 variables (unit)</i>			
Regime (ref.: Remand prison)			
Prison unit	0.32	0.11	0.00
Extra care unit	-0.81	0.30	0.01
Open unit	0.58	0.19	0.00
Addict unit	-0.13	0.43	0.00
Maximum security unit	0.68	0.31	0.03
Psychiatric unit	0.82	0.13	0.00
Women unit	0.53	0.25	0.04
Staff's tenure (years)	0.00	0.01	0.97
<i>Level 2 variables (prison)</i>			
Inmates' average age	NA		
Average time served	NA		
Proportion violent offenders	NA		
Building (ref.: Wing / Cruciform)			
(Stacked) Pavilion building	-0.64	0.32	0.05
Panopticon	0.86	0.33	0.01
Double bunking	NA		
Staff prisoner ratio	2.44	0.89	0.01
Constant	-3.22	0.46	0.00

We also run a multilevel regression model for the survey scale measure of *staff's feelings of safety*. Table 2 shows the coefficients that indicate the connection between ND factors and the performance scores. Only one prison level variable reaches the 0.10 alpha level, namely 'double bunking.' The other prison level factors are therefore eliminated.

Table 2. Multilevel linear regression model for dependent variable *staff's feelings of safety*. N=1,689 officers in 172 units in 43 prisons. Wald $\chi^2(9) = 24.62$. Prob = 0.00. NA = not applicable.

Staff's feelings of safety	B	S.E.	Sign
<i>Level 2 variables</i>			
Cell capacity of prison unit	0.00	0.00	0.02
Regime (ref.: Remand prison)			
Prison unit	-0.05	0.06	0.43
Extra care unit	0.22	0.11	0.04
Open unit	0.11	0.10	0.24
Addict unit	-0.23	0.13	0.09
Maximum security unit	-0.21	0.19	0.26
Psychiatric unit	0.14	0.10	0.18
Women unit	0.07	0.13	0.60
<i>Level 3 variables</i>			
Inmates' average age	NA		
Average time served	NA		
Proportion violent offenders	NA		
Building (ref.: Wing / Cruciform)	NA		
(Stacked) Pavilion building	NA		
Panopticon	NA		
Double bunking	.02	.01	.05
Staff prisoner ratio	NA		
Constant	3.12	0.11	0.00

The variables that reach significance in the model are cell capacity, regime and double bunking. In Step 1 we detected that 7.1% of the variance is attributed to differences between prisons (this is the model without explanatory variables). With the inclusion of the factors in the model above, 6.3% of the variance between prisons remains unexplained. Thus, the ND factors explain 0.8% of the variance on the prison level. We conclude that in this performance measure ND factors are of influence as well, but there is also sufficient variance (namely 6.3% of the total variance) that represents the effort of the prison.

5. Step 4: Interpretation

The presentation and interpretation of performance comparisons require substantial attention in order to bring the performance comparisons to a successful ending. Thorough consideration of this final step can prevent pitfalls such as misuse and perverse effects.

Using the regression models of Step 3, for every organization an empirical Bayes residual can be derived. This can be done in several statistical programs. We use the *reffects* function in Stata to estimate empirical Bayes residuals. The residuals on the organizational level are equivalent to the estimated (random) intercepts for the organizations in the comparison (Hox, 2010). These residuals reflect the valid variance (Tsai & Bridges, 2011), i.e., performance. The use of empirical Bayes residuals for performance comparisons improves the comparability between organizations because the influence of ND factors is eliminated. The interpretation of the residuals in terms of performance is quite unambiguous: inferior performers have a negative residual, superior performers have a positive residual (or the other way around when the tool measures an undesirable phenomenon, like *violent incidents*). Empirical Bayes residuals have the advantage that their precision is known. When confidence intervals are calculated as well, the significance of the mutual deviation between organizations can easily be seen.

For interpretational purposes, residual projections like caterpillar plots have at least one disadvantage, namely they do not reveal the underlying observed scores (Hood, Dixon & Beeston, 2008; Barnow & Heinrich, 2010). Our practical experience is that working floor managers and administrators want to keep an eye on the observed scores. Many people find it difficult to understand and interpret residuals or adjusted scores. Furthermore, the performance score of an organization might not lead to an extreme value of the residual, and the *observed* score might be divergent in such a way that higher level management or central executing agencies want to interfere. Therefore, it is desirable to project the original performance score next to the residuals.

5.1 Illustrations for Step 4

The residuals of our prison performance measures are placed in caterpillar plots (Figures 2 and 3). The point in the middle of every interval is the value of the residual on which the rankings are based. If the interval for a prison includes 0, this prison does not deviate significantly from the mean residual. Roughly, if the interval for one prison does not overlap with the interval of another prison (via visual inspection), the prisons differ significantly on the particular performance measure (a correct test is based on the variance of the difference between the two residuals, which is less conservative). The best performing prisons are found on the left-hand side of the plots, and the poorest on the right-hand side. For example, the most left prison in Figure 3 has a higher score (over 0.4) than might be expected given the non-discretionary factors that apply to this organization.

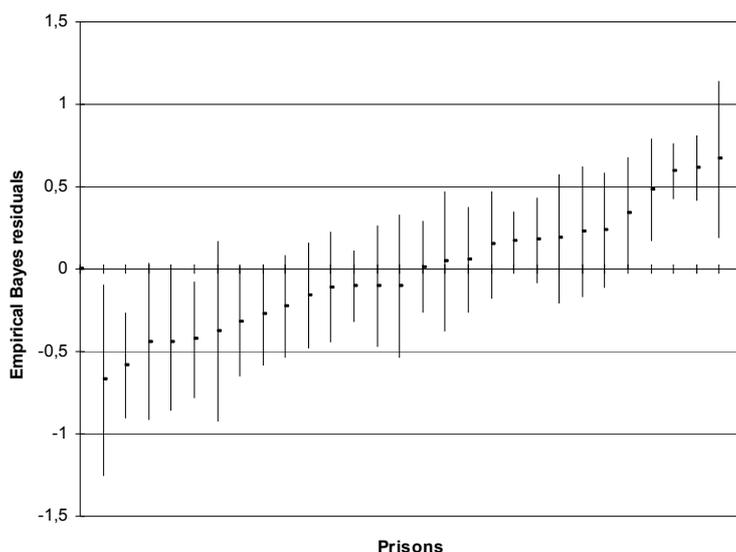


Figure 2. Caterpillar plot for the performance measures violent incidents. Confidence intervals set at 95%.

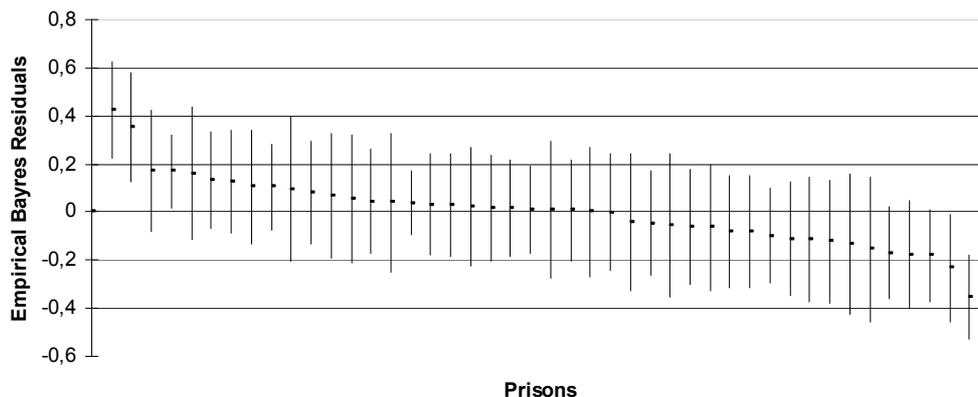


Figure 3. Caterpillar plot for the performance measures staff's feelings of safety. Confidence intervals set at 95%.

Below, Figures 4 and 5 simultaneously display observed scores (on the vertical axes on the left) and residuals (on the vertical axes on the right) for the two performance measures. In this way, one makes sound and fair performance comparisons on the one hand (by accounting for ND factors and determining the ranking of the residuals) while on the other hand keeping an eye on the real situation (observed score).

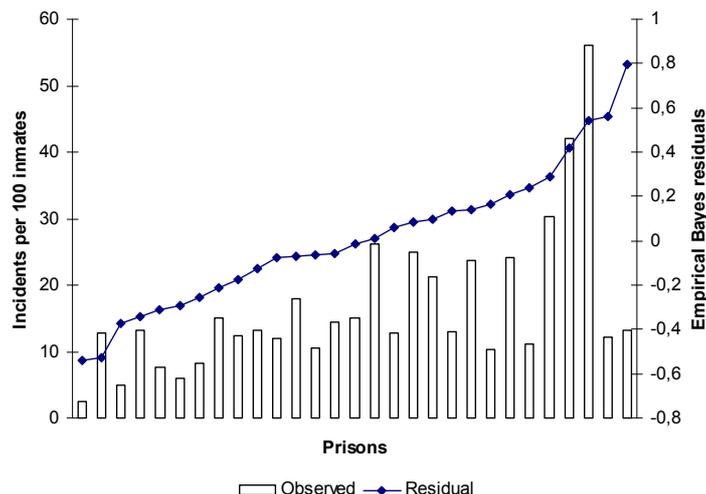


Figure 4. Observed scores and residuals for violent incidents. Prisons are ranked using their empirical Bayes residuals.

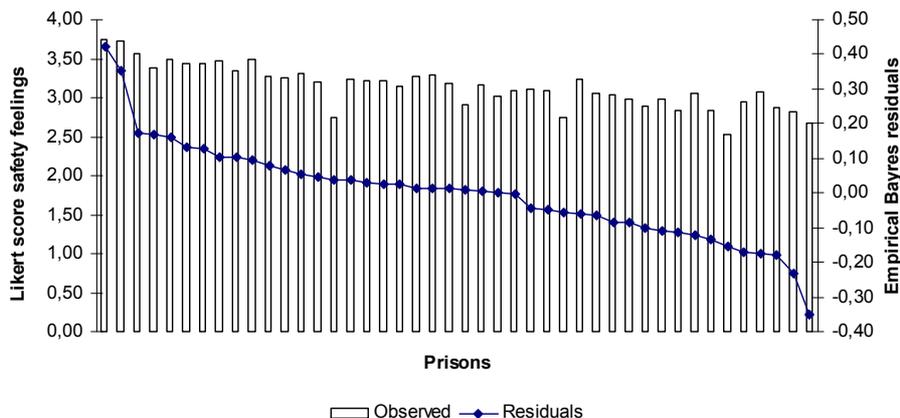


Figure 5. Observed scores and residuals for staff’s feelings of safety. Prisons are ranked using their empirical Bayes residuals.

We investigated whether rankings of prisons differ when these rankings are based on either observed scores or residual scores. The maximum change in rankings for the *violent incidents* measure was 18 ranks between the observed and the residual ranking. Only 1 out of 29 prisons kept the same rank. The rankings of *staff’s feelings of safety* changed up to 26 ranks with only 7 out of 43 prisons remaining in the same position rank. Therefore, we conclude that the ranking of adjusted performance measures differs substantially from the ranking of observed (i.e., unadjusted) performance measures.

With projections as presented in Figures 4 and 5, the performance analyst can easily assess the best performing prisons that are on the far left, while the prisons on the far right of the figures are performing worst. Strong and poor performers might undertake benchmark activities (e.g., exchange good practices) for improving the perceived safety of prison staff and reducing violent incidents. Furthermore, it is easy to detect some remarkable moderate performers in the middle of the figures (especially Figure 4). These prisons do perform as expected (residual around zero) but have divergent observed scores. These divergent scores seem to originate from factors out of the sphere of influence of prison management.

6. Closing Remarks

In this contribution we proposed a step-by-step plan for fair and sound performance comparisons between organizations. After the assessment of reliability, validity and variability of performance measures, the plan prescribes to examine factors that influence the performance measure. A performance measure can only play a useful role in performance comparisons if the variance in a performance measure is (partly) associated with the efforts of the management level under evaluation (the performing actors in the comparison). The performance measures must be adjusted for non-discretionary (ND) factors, these are factors that are connected to the performance measures but cannot be influenced by management. These adjustments can be made with the use of multilevel regression models. The next step is to rank the organizations on the adjusted scores, for which so-called empirical Bayes residuals are used. We recommend displaying the observed score for easy interpretation and as a 'warning function' for higher-level management or central executing agencies (or even politicians).

The main result is that the step-by-step plan 'levels the playing field' and therefore promotes fair and sound performance comparisons. A consequence might be that the application of the plan produces 'more valuable information for policy makers to use for both program management and accountability purposes' (Barnow & Heinrich, 2010: p. 66). Furthermore, there is an increased chance that performance measurement detects the right benchmark organization and exchange of best practices will indeed lead to performance improvement.

The way performance is measured always needs critical examination. Therefore, adjustment for measurement error is also considered. In this study, since seasoned staff might record *violent incidents* more accurately, we adjusted for tenure of staff. In the literature, survey measures are not always trusted (Camp, 1999). However, we assume that the staff survey measure is not seriously threatened by this specific measurement bias since the data are collected for the purpose of organizational improvement and derived from a representative sample. Nevertheless, it may be argued that measurement bias in survey data might arise because staff or customers of a particular organization exaggerates how difficult their (working) conditions are to incite management to take certain actions (or for other reasons). Next to this form of 'impression management' of respondents, another criticism to the use of survey data is that respondents of surveys would be self-deceptive in their answers (Paulhus, 1984). This means that respondents would report a state of affairs while knowing reality is different. However, other research suggests that survey results in prisons (both inmate and staff) vary in a systematic way across facilities (Camp, 1999; Camp et al., 2002; Molleman, 2008; Molleman & Leeuw, 2012). Survey data have been shown to be consistent with official prison records (Dagget & Camp, 2009; Molleman, 2011). Thus, the differences between prisons that we are able to detect with survey data seem to be valid and do not appear to be biased by impression management and self-deception.

The presented step-by-step plan of performance analysis and comparison has important practical benefits. When the methodology is well adapted, a manager cannot use excuses for poor performance concerning the non-discretionary factors (since the performance scores are adjusted for these factors). A related benefit is that the methodology prevents for 'cream skinning' or 'cherry picking' because it pushes back the incentive to search for easy circumstances.

Although we believe that the step-by-step plan promotes fair and sound performance comparisons, the risk of a rigid interpretation of the outcome of the analysis when only numbers and figures are involved still remains. The interpretation of performance measures should be accompanied with an open discussion between managers and their superiors (Halachmi, 2005). The authors agree with Barnow and Heinrich (2010) that statistical modeling should be viewed as a complement rather than a substitute for negotiating performance standards. In consultations on performance assessment where managers give account, at least three stages should be considered:

- The development of performance scores in the course of time in the specific organization;
- A contrast between the scores which are observed and the goals of which the manager and his or her principal agreed on beforehand and;
- A coherent and integral analysis of several measures with clarification concerning content by the manager (to make 'the story behind the numbers' more explicit). A moderate score on a measure such as *staff's feelings of safety* may be caused by structural shortcomings in rule enforcement and continuous poor backing of colleagues on the working floor. In a discussion on performance between the manager and the principle, a totally different explanation may also come up. A moderate score on *staff's feeling of safety* can be found in organizations that are very safe in the ordinary course of events. Due to a single horrible incident, the *feelings of staff's safety* may suddenly reach rock bottom. It is clear that further discussion of performance figures is needed.

The ND factors selected may not be influenced by the management level that is subject to performance comparison; nonetheless, these factors may be influenced by higher (or even lower) level managers. According to Stiefel et al., (1999) some factors might be controllable at one specific level of an organization but uncontrollable at another. Although an organization might have, for example, very unfavorable fixed conditions (e.g., mediocre accommodation and a difficult clientele) and its poor observed performance score do not therefore make the organization a poor performer, the situation might exceed acceptable limits. If it comes to that, it is desirable that higher-level management receives a signal.

References

- Barnow, B. S., & Heinrich, C. J. (2010). One standard fits all? The pros and cons of performance standard adjustments. *Public Administration Review*, 70(1), 60-71. <http://dx.doi.org/10.1111/j.1540-6210.2009.02111.x>
- Camanho, A. S., Portela, M. C., & Vaz, C. B. (2009). Efficiency analysis accounting for internal and external non-discretionary factors. *Computers & Operations Research*, 36(5), 1591-1601.
- Camp, R. C. (1989). *Benchmarking: The Search for Industry best practices that lead to superior performance*. ASQ Quality Press: Milwaukee. <http://dx.doi.org/10.1016/j.cor.2008.03.002>
- Camp, S. D. (1999). Do inmate survey data reflect prison conditions? Using surveys to assess prison conditions of confinement. *The Prison Journal*, 79(2), 250-268. <http://dx.doi.org/10.1177/0032885599079002007>
- Camp, S. D., Gaes, G. G., Klein-Saffran, J., Daggett, D. M., & Saylor, W. G. (2002). Using inmate survey data in assessing prison performance: A case study comparing private and public prisons. *Criminal Justice Review*, 27(1), 26-51. <http://dx.doi.org/10.1177/073401680202700103>
- Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2, 429-444. [http://dx.doi.org/10.1016/0377-2217\(78\)90138-8](http://dx.doi.org/10.1016/0377-2217(78)90138-8)
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302. <http://dx.doi.org/10.1037/h0040957>
- Cronbach, L. J. (1947). Test reliability: Its meaning and determination. *Psychometrika*, 12, 1-16. <http://dx.doi.org/10.1007/BF02289289>
- Daggett, D. M., & Camp, S. D. (2010). Do official misconduct data tell the same story as the individuals who live in prison?. *Criminal Justice Review*, 35, 200-219. <http://dx.doi.org/10.1177/0734016808329291>
- Drösler, S. E., Romano, P. S., Tancredi, D. J., & Klazinga, N. S. (2012). International Comparability of Patient Safety Indicators in 15 OECD Member Countries: A Methodological Approach of Adjustment by Secondary Diagnoses. *Health Service Research*, 47(1), 275-292. <http://dx.doi.org/10.1111/j.1475-6773.2011.01290.x>
- Fried, H. O., Lovell, C. A. K., Schmidt, S. S., & Yaisawarnng, S. (2002). Accounting for environmental effects and statistical noise in Data Envelopment Analysis. *Journal of Productivity Analysis*, 17, 157-174. <http://dx.doi.org/10.1023/A:1013548723393>
- Fung, V., Schmittdiel, J. A., Fireman, B., Meer, A., Thomas, S., Smider, N., Hsu, J., & Selby, J. (2010). Meaningful variation in performance: a systematic literature review. *Medical Care*, 48, 140-148.
- Gaes, G. G., Camp, S. D., Nelson, J. B., & Saylor, W. G. (2004). *Measuring Prison Performance, government privatization & accountability*. California: AltaMira Press. <http://dx.doi.org/10.1097/MLR.0b013e3181bd4dc3>
- Halachmi, A. (2005). Performance measurement is only one way of managing performance. *International Journal of Productivity and Performance Management*, 54(7), 502-516. <http://dx.doi.org/10.1108/17410400510622197>
- 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://dx.doi.org/10.1080/10967490802301286>
- Hox, J. J. (2010). *Multilevel Analysis: Techniques and applications* (2nd ed.). New York: Routledge.
- Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come Bayesian Methods for Data Analysis in the Organizational Sciences. *Organizational Research Methods*, 15(4), 722-752. <http://dx.doi.org/10.1177/1094428112457829>

- Laird, N. M., & Louis, T. A. (1989). Empirical Bayes ranking methods. *Journal of Educational Statistics*, 14(1), 29-46. <http://dx.doi.org/10.2307/1164724>
- Landon, B., Iezzoni, L. I., Ash, A. S., Shwartz, M., Daley, J., Hughes, J. S., & Mackiernan, Y. D. (1996). Judging Hospitals by Severity-Adjusted Mortality Rates: The Case of CABG Surgery. *Inquiry*, 33(2), 155-166.
- Leyland, A. H., & Groenewegen, P. P. (2003). Multilevel modelling and public health policy. *Scandinavian Journal of Public Health*, 31, 267-274. <http://dx.doi.org/10.1080/14034940210165028>
- March, J. G., & Sutton, R. I. (1997). Crossroads - Organisational Performance as a Dependent Variable. *Organization Science*, 8, 698-706. <http://dx.doi.org/10.1287/orsc.8.6.698>
- Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, 16(3), 283-301. [http://dx.doi.org/10.1016/S0272-7757\(96\)00081-7](http://dx.doi.org/10.1016/S0272-7757(96)00081-7)
- Molleman, T., & Leeuw, F. L. (2012). The influence of prison staff on inmate conditions: A multilevel approach to staff and inmate surveys. *European Journal on Criminal Policy and Research*, 18(2), 217-233. <http://dx.doi.org/10.1007/s10610-011-9158-7>
- Molleman, T. (2008). *Psychometrische Kwaliteit van en de verbanden tussen de gedetineerdensurvey en de BASAM-DJI*. Den Haag: WODC.
- Molleman, T. (2011). *Benchmarking in the prison system: A study on the possibilities of comparing and improving performance*. Den Haag: Boom Juridische Uitgevers.
- Nyhan, R. C., & Martin, L. L. (1999). Comparative performance measurement: A primer on data envelopment analysis. *Public Productivity & Management*, 22(3), 348-364. <http://dx.doi.org/10.2307/3380708>
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598-609. <http://dx.doi.org/10.1037/0022-3514.46.3.598>
- Pidd, M. (2012). *Measuring the performance of public services*. Cambridge: Cambridge University Press.
- Poister, T. H. (2010). Performance Measurement: Monitoring Program Outcomes. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of Practical Program Evaluation* (3rd ed., pp. 100-124). San Francisco: Jossey-Bass.
- Radnor, Z. J., & Barnes, D. (2007). Historical analysis of performance measurement and management in operations management. *International Journal of Productivity and Performance Management*, 56(5/6), 384-396. <http://dx.doi.org/10.1108/17410400710757105>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models* (2nd ed.). Thousand Oaks: Sage.
- Silber, J., Rosenbaum, P. R., & Ross, R. N. (1995). Comparing the contributions of groups of predictors: which outcomes vary with hospital rather than patient characteristics?. *Journal of the American Statistical Association*, 90, 7-18. <http://dx.doi.org/10.2307/2291124>
- Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration*, 18(2/3), 277-310. <http://dx.doi.org/10.1080/01900699508525011>
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London: Sage.
- Staiger, D. O., Dimick, J. B., Baser, O., Fan, Z., & Birkmeyer, J. D. (2009). Empirically derived composite measures of surgical performance. *Medical Care*, 47, 226-233. <http://dx.doi.org/10.1097/MLR.0b013e3181847574>
- Stiefel, L., Rubenstein, R., & Schwartz, A. E. (1999). Using Adjusted Performance Measures for Evaluating Resource Use. *Public Budgeting and Finance*, 19(3), 67-87. <http://dx.doi.org/10.1046/j.0275-1100.1999.01172.x>
- Tsai, A., & Bridges, J. F. P. (2011). Statistical and Econometric Risk Adjustment Methods for Measuring the Quality of Hospitals. *Journal of Health Policy, Insurance, and Management*, 1, 45-61.

Appendix

Staff's feelings of safety (Cronbach's α is 0.86). Items are 5-point Likert scales, ranging from 'totally disagree' to 'totally agree'.

1. The working environment has been designed to make me feel safe.
2. Everything possible is done here to guarantee my safety.
3. The work has been organized in such a way that nothing serious can happen to me.
4. I feel at ease when I walk through the building.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).