Performance measurement in the Dutch prison system

Methodological guidance for public sector performance assessment

Toon Molleman

Printing Ipskamp drukkers, Enschede

Cover design Eefje Ossevoort

ISBN 978-94-6259-232-2

Performance measurement in the Dutch prison system

Methodological guidance for public sector performance assessment

Prestatiemeting in het Nederlandse gevangeniswezen

Methodologische handvatten voor het meten van de prestaties van organisaties in de publieke sector (met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 4 juli 2014 des middags te 12.45 uur

door

Toon Molleman

geboren op 19 oktober 1981

te Nijmegen

Promotors: Prof. dr. P.G.M. van der Heijden

Prof. dr. F.L. Leeuw

This book was made possible by the financial support of the Dutch Ministry of Security and Justice, in particular the Directorate-General of Prevention and Sanctions and the Custodial Institutions Agency (in Dutch: Dienst Justitiële Inrichtingen, DJI). This support included the availability of the data and the access to prisons and interviewees. The study was conducted at the Documentation and Research Centre (in Dutch: Wetenschappelijk Onderzoeks- en documentatiecentrum, WODC) of the Ministry of Security and Justice. This publication is financially supported by WODC. Previously, a WODC research report came out on the topic of this book (Molleman, 2011a). The author declares that there are no conflicts of interest.

Manuscript committee: Prof. dr. K. van den Bos (Utrecht University)

Prof. dr. P. Nieuwbeerta (Leiden University)

Prof. dr. J.J.C.M. Hox (Utrecht University)

Prof. dr. H.G. van de Bunt (Erasmus University Rotterdam)

Prof. dr. H. de Groot (University of Twente)

“Not everything that counts can be counted, and not everything that can be counted counts” (Cameron, 1963: 13)

Contents

Chapter 3

Chapter 4

Chapter 1

Introduction

1.1 Background

During the last three decades performance measurement became an important instrument in public and semi-public organizations (Light, 2006; Hood, 1995; Pollitt & Bouckaert, 2011; Pidd, 2012). One reason for its rise in importance was that civil services were placed 'at arm's length' of central governments in many western countries. By increasing their autonomy and using more entrepreneurial management instruments,[1] (semi-)public organizations were believed to become more adaptable, responsive, efficient and effective (Osborne & Gaebler, 1992). Deregulation and the 'let managers manage' philosophy were to a greater or lesser extent applied to hospitals, schools, universities, prisons and other organizations in the (semi-)public sector. Simultaneously, central governments wanted to safeguard for (semi-)public sector organizations to indeed realize their tasks and goals in an efficient way. Therefore, instruments for performance measurement[2] were installed in order to collect data on, for instance, effectiveness and efficiency. Talbot (2010) shows that performance measurement of (semi-)public services has been on the rise in many countries over the last decades, while Lonsdale (2011) notes that many countries have laid down mandatory accountability in regulations and legislation.

Performance measurement is not only thought to be relevant for reasons of accountability, it is also expected to contribute to the improvement of performance. By comparing and contrasting performance scores of different organizations, it is believed that managers and employees are motivated to work harder and smarter (Poister, 2010a; Behn, 2003). Van Loocke and Put (2011) found evidence that performance measurement (and performance audits in particular) do have these consequences (other pieces of evidence are found in for example: Bevan, 2009; Propper, Sutton, Whitnal & Windmeijer, 2010; Bloom, Propper, Seiler & Van Reenen, 2010). Despite the fact that the impact of performance measurement on performance itself was never studied in a well-structured manner, Poister (2010b) concludes that fragmented evidence in several (semi-)public sectors suggests that performance measurement contributes to the

1 In the course of time, different management philosophies came into fashion among which New Public Management probably received most attention.

2 There are a variety of performance instruments, such as Planning and Control Mechanisms and Accountability Contracts, that include measurement activities.

improvement of performance. The conclusion seems justified that measuring performance can make a difference, however 'not always or as often as would be desirable' (Lonsdale, Ling & Wilkins, 2011: 321).

Despite the advantages of performance measurement, unexpected and even negative side effects also became evident (e.g. Smith, 1995; Van Thiel & Leeuw, 2002; Pidd, 2005; De Wolf & Janssens, 2007; Pollitt & Bouckaert, 2011; Pollitt, 2013). For instance, the utilization of performance measurement made local administrators now and then 'game the numbers', 'cook the books' and 'learn perverse' in favor of –deliberately or unintentionally– 'looking good' for central government. Evidence also shows that ignorance of potential side effects may deteriorate performance (Smith, 1995). Several manifestations of undesirable side effects of performance measurement are addressed throughout this book.

To sum up, evidence exists – although fragmented – for the effectiveness of performance measurement; at the same time, the risk of negative side effects is shown to be present. In the present study we draw lessons from these general findings by focusing on *conditions* capable of contributing to accountability and performance improvement while preventing negative side effects. We will do that for a specific sector, namely the Dutch prison system.

The measurement of performance starts with a clear understanding of the goals organizations want to realize. However, organizational goals are sometimes only expressed in somewhat conceptual or even vague 'mission statements'. This is especially true when (semi-)public organizations have complex missions that can be perceived differently by various relevant actors; a thorough analysis is required before performance can be measured (Behn, 2003).

A second condition is to ensure that measuring organizational performance takes place in a valid and reliable way. Achieving this, however, is easier said than done. The problem is that every single data collection method may be biased and may create side effects. As early as 1934, LaPiere found evidence that what people say (or write) is not always congruent with what they do (LaPiere, 1934). Obviously, this and similar types of

discrepancies also play a role, to a greater or a lesser extent, when measuring performance. A solution to this problem is the use of *multiple data collection methods* to assess performance. By using multiple methods simultaneously, weaknesses of certain methods may be counterbalanced by strengths of others. Therefore, an explicit measurement strategy is needed if and when one strives for valid and reliable data on organizational performance (Kravchuk & Schack, 1996).

A third condition for an effective performance measurement concerns knowledge about *how* an organization can improve its performance. Admittedly, measuring performance is important as it addresses the criterion of accountability; however, learning from these measurements in order to improve performance is also important. Therefore, analyzing the determinants of performance may contribute to evidence-based managerial actions to improve performance. Part of this analysis is the unpacking of who (and with which means) is capable of influencing performance of an organization.

As is well-known, the management of an organization has decision-making powers, but they may not get grip on every factor that is related to organizational performance. Organizations have to perform within a context that is partly not malleable by the management. However, these 'given circumstances' may bear a large share in organizational performance (Gaes, Camp, Nelson & Saylor, 2004; Talbot, 2010).[3] In the context of prisons, for example, the characteristics of the inmate population in terms of age and ethnical background may differ between prisons and may seriously co-determine the extent to which violence appears. Since local prison managers, at least in the Netherlands, do not decide which inmates are included or excluded in the population of their prison, 'raw' scores of violent incidents would not be appropriate to assess the managerial performance because they also reflect the given circumstances. Methodological and statistical techniques may ensure doing 'apples to apples comparisons' by ruling out factors that are not changeable by management of prisons. If given circumstances are taken into account and knowledge is available about how

3 Performance measures may not be fully the result of managerial efforts since the measurement is 'inserted in systems which are already fluid and changing' (Pawson, 2013: 5). Beforehand it is not known whether changes in performance are the effect of organizational decisions by the management or by others, the effect of measurement activities and their unintended (side) effects and/or represent changes that would have happened anyway.

performance can be changed by the management, very probably effective performance measurement is more in reach.

These conditions are expected to contribute to accountability of organizations, performance improvement and the reduction of negative side effects; however, the added value of the conditions are supported by only fragmented evidence and are tested in only a limited number of branches of the (semi-)public service. Therefore, it would be too optimistic to believe that performance measurement guarantees a self-regulating organization after implementing the mentioned conditions. Therefore an ex-ante assessment should be carried out to ask stakeholders (like managers and policy makers) how they assess the chance that performance measurement, given the studied conditions, indeed will lead to accountability and performance improvement.

The present study focuses on how these conditions may apply to the prison system of the Netherlands; this system and its performance instruments are further described in the next paragraph.

1.2 Empirical subject of the study: the Dutch prison system

Prisons are found anywhere in the world. In many countries the prison system strives to reach the goals of safe and humane prison conditions, preventing re-offending and promoting the re-integration of inmates into society. Prisons are thought to be extraordinary (semi-)public services because they have physically closed buildings with staff and inmate cultures separated from the world outside the walls. Furthermore, prisons are potentially violent places because of the criminal propensity of the inmate population. Inmates are limited in their freedom of movement and are held to the house rules and the daily schedule. Although these elements make prisons special, prisons share characteristics with organizations that are labeled *total institutions* in sociology.[4] Residents in total institutions face similar conditions and are to a certain extent limited in their external social contacts (Goffman, 1961). Moreover, the opportunities to leave the institution are limited and the care for the residents is in the hands of a (semi–)

4 Examples are psychiatric clinics, orphanages, homes for the elderly, army bases and boarding schools.

public organization.[5] On this point, the findings in the present study may have a broader application than to Dutch prisons alone. That is, prisons have somewhat similar operational management issues, as does any total institution, such as a financial administration.

The prison system in the Netherlands operates as an executive agency (officially named Custodial Institutions Agency) and serves under the responsibility of the Ministry of Security and Justice. The agency has its headquarters in The Hague where the national prison management resides. Prisons managers across the Netherlands have considerable managerial autonomy (established in the Dutch Penitentiary Act). Penal legislation stipulates that the manager of each prison has decisive powers regarding how to implement and apply the regimen and the house rules in daily prison life. Seen from the national prison management, prisons are placed 'at arm's length' which promotes the 'let the manager manage' principle. Since the Ministry is responsible for all Dutch prisons, performance measurement is an important source of information in order to monitor whether or not the goals are realized in an efficient way. Therefore, prisons keep a variety of databases (up to date). Examples of those registrations are reports of violent incidents and drug monitoring, staff's performance interviews and sickness absence, complaints by inmates, realization of educational programs, aftercare activities and awarded inmate furloughs. In addition to these registrations, the agency is subject of reports on capacity and financial administration, (security) audits, supervision by inspectorates and surveys among staff and inmates.

In this study, data from different sources in the fiscal years 2006-2007 are used. Although these data may look outdated, for the purpose of this book that is not a problem, since all of the databases used are still operated to date and the focus in this book is on methodological developments and not on performance judgments in practice. The data are collected at the national level as well as in the 45 prisons that were operating in the Netherlands during the study period.

5 A total institution exercises total control over its population. 'Every movement is controlled by the institution's staff; an entirely separate social world comes into existence within the institution, which defines the inmate's social status, his relationship to all others, his very identity as a person' (Wallace, 1971: 1-2).

Many of these data sources are input for the performance measurement of the Dutch prison system. With those measurements, performance comparisons between prisons are drawn up for accountability reasons. A second, and probably just as important, application of the measurements are activities concerning improvement of performance. Comparing the performance of prisons (and prison units) may lead to benchmark activities.

The national prison management of the executive agency and several local prison managers have expressed their interest in the present study. They expect that the methodological guidance given in this study might lead to better opportunities for accountability and performance improvement. Moreover, there seems to be space for performance improvement in particular prisons since there is evidence that the performance scores vary between Dutch prisons (Boin, 2001; Inspectorate for the Implementation of Sanctions, 2009; Molleman, 2011a). The mechanism of comparing and contrasting performance may stimulate prisons that perform on relatively lower level to start similar efforts as better performing prisons have already used. Having discussed the research issue, the research subject and the data sources, we now get to the research questions that will be investigated in this book.

1.3 Research questions

In the first paragraph of this chapter, several conditions of performance measurement are described. We know apply these conditions to the Dutch prison system in order to search for an accurate instrument for performance measurement. The central research question of this book is as follows:

How may conditions for accurate measurement of performance be applied to the Dutch prison system and may performance measurement within those conditions lead to an increased likelihood of accountability and performance improvement in Dutch prisons?

The conditions addressed in the first paragraph of this chapter may contribute to an accurate instrument for performance measurement in the Dutch prison system. A first

condition is the specification of the organizational goals. This leads to the first sub-question:

- Following the goals of imprisonment, what are the tasks of Dutch prisons?

When the tasks are well described, certain measurements may assess the performance on these tasks. However, comparing organizations on these measurement results can be unfair. Organizations may have different contextual circumstances that they cannot change. In case these differences affect performance scores, the comparison of 'raw scores' seems not appropriate. Before we compare performance between organizations, and that is the second condition of performance measurement, we should account for contextual circumstances. This leads to the second sub-question:

- Do Dutch prisons have equal contextual circumstances, and if they do not, how can these differences be dealt with when making performance comparisons?

If performance assessment relies on just a single measurement, we run the risk that the measurement is biased and the assessment generates a distorted view on organizational performance. Furthermore, organizational performance of (semi-)public services, and prisons in particular, is sometimes multifaceted; therefore, multiple measurements methods are sometimes considered. To date, we lack guidance on which measurement methods may be included and how these can be combined to assess the multifaceted character of organizational performance. Therefore, a third condition of performance measurement is to have a measurement strategy that guides the assessment of performance of Dutch prisons.

- Which measurement strategy may comprehensively assess organizational performance, and give an account of limitations of various data collection methods?

Ideally, the differences in performance scores between organizations is not affected by contextual circumstances they *cannot* influence; as posed in the second sub-question. The key to fair performance measurement is to compare prisons on performance score differences that result from factors that *can* be influenced by a local prison management.

Knowledge about those factors may help the organizational management to improve organizational performance. The next sub-question therefore is:

- Do performance measurements relate to factors that can be influenced by a local prison management?

Finally, we want to know more about the likelihood that performance measurement in the Dutch prison system, in case the above conditions are applied, indeed contributes to accountability and performance improvement.

- According to relevant stakeholders, will the findings in this study contribute to accountability and performance improvement in the Dutch prison system?

1.4 Central definitions

A central expression in this book is *performance*. This refers to the extent to which an organization succeeds to accomplish its goals. *Performance management* includes all managerial activity at any level in an organization that affects performance. For example, this can be done by comparing individual units within an organization in order to exchange good practices or by generating (pseudo-)competition (Propper & Wilson, 2003).

Terms like (data) *sources*, *data collection methods* and *indicators* are used frequently in this book. These are all elements of the approximation of performance and are considered crucial to performance management (Fryer, Antony & Ogden, 2009). Data sources are places where figures and other information can be found concerning organizational activities. Data collection methods refer to specific methods used for the measurement of a certain phenomenon; for example, the performance or goal accomplishment of an organization. A performance indicator is an approximation of (certain aspects of) performance in reality in numerical terms. Performance indicators often include a comparison between the achieved score and some norm or (national) average presented in a comparative graph or ranking.

1.5 The role of theories

The respective research questions of the present study are investigated with the guidance of theories. This theory-based approach makes it possible to learn and benefit from earlier insights in the literature. The theories that we have applied are briefly described.

In order to qualify the areas in which the Dutch prison system has to perform – which is the subject of the first research sub-question (see paragraph 1.3)– *penological theory* is used in an analysis of the reasons for a judge to issue a prison sentence (Franke, 1995). With regard to the second research question, we refer to *benchmarking theory* to make the mechanism behind practices of performance comparisons explicit (Poister, 2010b; Meyer, 1997). To answer the third research question concerning the measurement strategy of organizational performance assessments, the theoretical assumption is adopted that different relevant actors may assess different parts of reality in different ways. In the present study we take the ontological standpoint that an independent reality exists, but that there are epistemological problems assessing it (Bhaskar, 2008). Critical realism (a philosophy of science) allows for multiple data collection methods to assess socially complex phenomena like organizational performance (Talbot, 2010). Since (semi-)public performance – and prison performance in particular – is multifaceted, this seems an appropriate stance. To give an answer to the fourth research question, we use import theory and deprivation theory (Sykes, 1958; Irwin & Cressey, 1962). These theories are tested to find clues about which part of the performance score variance between prisons is due to managerial efforts.

1.6 Structure of the book

The next four chapters deal with the research questions as stated in paragraph 1.3. In the sixth chapter of this book an assessment made by stakeholders will be presented about what they expect of a performance instrument in the Dutch prison system using the conditions of performance measurement discussed in this book. In the final chapter, the research questions are answered, the limitations of the study are mentioned as well as the possibilities for broader application of the results.

Chapter 2

A matter of balance: on the abstract goals of imprisonment and task ambiguity in Dutch prisons

This chapter previously appeared as: T. Molleman & A.A. van den Hurk (2012). Een kwestie van evenwichtskunst: Over de doelen en taken van het gevangeniswezen, *Delikt & Delinkwent*, 55(7), 576-590. An English version is currently under review.

Abstract

According to national and international prison inspectorates, considerable differences are found in the implementation of prison sentences within countries. Laws and regulations should guide prisons to realize uniformity of implementation so as to secure equal prison conditions for inmates across prisons as well as to secure the interests of victims and the broader society. These rules may be clear when it comes to the key goals of imprisonment, since most prisons should provide safe, humane and rehabilitative conditions. But if we look more closely, complex ambiguities seem to be at work. Next to the interests of inmates, victims and society may be at issue, these ambiguities may create a situation in which prison organizations themselves do not know exactly for what they are deemed responsible. In this chapter the specific situation of the Dutch prison system is analyzed to show how implementation differences can arise within a seemingly detailed regulatory framework. The chapter aims to sum up the tasks of prisons and make the ambiguities in these tasks explicit.

2.1 Introduction

In many European countries and beyond, prisons strive for the safety of society, inmates and staff as well as humane conditions for and rehabilitation of inmates. Other than the fact that these topics are established in mission statements of the prison services, national and international laws give guidance for the implementation of prison sentences, such as the European Prison Rules (Council of Europe, 2006) and, in the case of the Netherlands, the Dutch Prison Act (Tweede Kamer, 1998). Furthermore, international minimum standards are supervised by the European Court of Human Rights (ECHR), the Committee of Prevention of Torture (CPT) and national inspectorates. Despite the mission statements, laws and inspectorates, it seems not to be totally clear how prisons should deliver their services.

The literature frequently refers to the goals of a judge when (s)he penalizes a criminal suspect. However, the elaboration of these goals into concrete tasks for correctional facilities is much less of a common subject of study. The absence of this elaboration is illustrated by the Dutch Inspectorate for the Implementation of Sanctions in 2009. The inspection found large differences in the applied procedures, making the 'quality of the implementation of sentences subject to local interpretation, arbitrariness and intuition' (Dutch Inspectorate for the Implementation of Sanctions, 2009: 35). To secure the interests of society, victims and inmates, a clear elaboration of tasks and operating procedures of detention is beneficial so as to reduce the chance of differences in the implementation of prison facilities. The facilities themselves can benefit from this exercise as well; putting the key objectives into practical tasks contributes to clarity on what exactly is demanded to be accomplished.

Every year, about 35,000 citizens face a prison sentence for varying periods of time (Molleman & Van den Hurk, 2012). These citizens are held in the penitentiaries of the Dutch Custodial Institutions Agency, hereafter called 'the prison system'. The prison system is associated with a wide range of goals, expectations and requirements concerning the implementation of the deprivation of liberty. Citizens (including victims) assume that inmates cannot escape from the facility and serve their time in not too spacious circumstances. Obviously, inmates and their families have different interests in

how the detention is implemented; examples are the possibility of visiting and the availability of amenities. The prison staff form a group of stakeholders as well. They want to work within safe conditions, find a challenge in their profession and be proud of their employment in the prison system. Finally, politicians have had a varying vision over the years on what the result of penal institutions should be. Sometimes retaliation prevails, sometimes social rehabilitation of inmates or reducing reoffending is the main objective. Clearly, the implementation of detention is a concern of various stakeholders who have partly conflicting interests and varying objectives.

To date, there is no analysis that bridges the gap between the legal basis of the deprivation of liberty and the concrete tasks of the workplace in prison. This chapter seeks to translate penal goals to concrete implementation tasks of prisons. It will be shown that translation is partly problematic because of uncertainty about the interpretation of the so-called principles of *minimal restrictions* and *rehabilitation*. On the one hand, the Dutch Prison Act stipulates that only a minimum of constraints should be imposed on inmates and a maximum effort be made to promote rehabilitation. On the other hand, however, there are rules in everyday practice such as restrictions on visits and the ban on (using) computers.

The elaboration of the goals of detention into concrete implementation tasks of prisons starts with an analysis of the history of the goals of the deprivation of liberty. Whether the themes of the goals have been consistent in recent history will be examined. Then a suggestion is presented of how to transform goals into implementation tasks by using the Dutch Prison Act and related regulations. This leads to a conceptual framework which shows where tasks and requirements of the implementation of the deprivation of liberty are *ambiguous*. The contribution concludes with some solutions to the aforementioned conflicting tasks in the deprivation of liberty in order to reduce vagueness about their interpretation.

2.2 A brief history of incarceration in the Netherlands

From the beginning of the nineteenth century, imprisonment became a topic of interest for Dutch administrators and scientists. The Enlightenment convinced society that

corporal and capital punishment were too cruel and ineffective. From that moment on, punishment focused on the psyche of the criminal. Besides the penalty of physical punishment, criminals had hitherto also been locked in 'reform houses' with overcrowded work halls and dormitories. Because of the disorderly situation in these accommodations and the fear of criminal contagion between inmates, the opportunities for social contact were restricted. However, additional suffering was no longer the main purpose of punishment; the use of solitary confinement was thought to bring about moral improvement (Scharff Smith, 2009). Silence and strict supervision were expected to suppress immoral behavior (Bentham, 1995). Moreover, solitary confinement was thought to have a deterrent effect on inmates. By incarcerating criminals it was furthermore intended to promote the security of society (the so-called incapacitation function).

Franke (1995) qualified the prevailing view of human beings as the *homo clausus*. This means that criminals are intrinsically bad and have a congenital talent to pursue evil; social and economic factors would have only a limited impact on crime. Years of solitary confinement would lead to religious awakening of his or her conscience and the decrease of immorality. This so-called Philadelphia system was maintained in the Netherlands until the end of the nineteenth century when administrators took account of the serious psychological effects of total isolation (such as the syndrome of Ganser or 'prison psychosis'). In the system, the health of inmates deteriorated and reoffending was not reduced.

The belief that detention should have a goal – and should not focus on the inmate's history – was born with *The New Direction* in criminal law, initiated by the German criminologist and legal expert Franz von Liszt. Certain groups of inmates were thought to have the potential to improve and should be entrusted to institutions for humane treatment ('Besserungsanstalten') that provided therapeutics such as labor and elementary education (Kempe, 1973). Crime was no longer seen as an exercise of one's free will and a trend was set to employ detention for the behavioral improvement of inmates (the 'pedagogical approach'). Because social factors were increasingly thought to be important, more social interaction was allowed in detention. This paradigm shift meant that the mechanism of behavioral change no longer relied on external coercion (deterrent conditions), but on internal coercion (Franke, 1995). At the start of the

twentieth century, internal compulsion was reflected in the introduction of suspended sentences and conditional release. Behavioral improvement was 'enforced' by the possibility of inmates earning personal benefits.

They left the cellular system with solitary confinement (total isolation) and the prison regime gradually allowed for more contact between inmates with solitary confinement at night (which is still more or less the situation for most inmates in the Netherlands today). The risk of criminal contamination was considered less harmful than the damage that solitary confinement entailed (e.g. insanity, self-mutilation and suicide). Retribution and deterrence were no longer targets that the prison system had to aim for, but rather the principles of legitimacy of the detention. The belief in enforced promises for the future of inmates gave way to educating inmates, providing a daily structure and practicing with freedoms within the prison walls. The humanization of prison conditions expanded step by step: differentiation and individualization became core concepts of the treatment and conditions of inmates. Amenities and privileges grew in size (such as receiving visitors and mail, opportunities for sports and exercise of religion) in which the interaction with staff was seen as crucial. A 'healthy' relationship with another human being would encourage personality change (Franke, 1995). Zwezerijnen (1972) noted that within the prison walls the 'command household' was slowly replaced by a 'negotiating household'. That change was also observed in the United States under the motto 'not coerce, but coax' (DiIulio, 1987: 19).

The 'healthy relationship' between prison staff and inmates was and is shaped in different ways. In the United States, some states primarily relied on informal contact with inmates. In other states the prison regime was relatively formal and the management strongly relied on discipline and rules (DiIulio, 1987). Although the contacts between inmates and staff have become more personal, the ways in which rehabilitation is implemented differs largely (Roth, 1985; Tewksbury & Mustaine, 2008). A similar state of affairs was also found in the Netherlands: the relationship between staff and inmates varies. In one prison, the relationship has the character of a "game" that is sometimes playful and then grim again. In another prison, there is hardly any relationship, because the contacts are minimal (Grapendaal, 1990).

The goals of rehabilitation and human dignity had a prominent place in the new Dutch penal legislation established in 1953. In the Netherlands and the further Western

world, the practice of punishment was driven by the ideal of rehabilitation until the seventies (Garland, 2001). Prison penalties were to rehabilitate and explicitly not to pursue any retributive goal. As a consequence, good prison conditions, treatment and no additional suffering were key elements of the prison regime (sometimes contained in the term *penal welfarism*). In the eighties, however, these principles changed.

The disappointing results (in terms of reoffending) and budget cutbacks led to retrenchment of the prison system in the last twenty years of the twentieth century. A government committee (on the 'psychiatric / therapeutic services of the prison system') reported in 1983 on the new vision of detention: 'former beliefs that undergoing the punishment would bring about rehabilitation through self-reflection and proper processing of feelings of guilt, cannot persist any longer facing the reality of hospitalization, increase of aggression, apathy, regression, moral decline, and alienation from family' (Franke, 1995: 752). In the policy document 'Role and Future' of 1982 the rehabilitation goal was still listed, but the possibility of influencing reoffending during detention were put into perspective (Hoekendijk & Kommer, 2011). In the detention plan of 1994 (named 'Efficient Detention') preventing harm to inmates was almost all that remained of the goals of improvement. In the new century the prison system again faced further cutbacks and austerity; certain policies were canceled or curtailed, such as differentiation (with the introduction of the standard regimen) and detention phasing (decrease of open and semi-open detention capacity). The rehabilitation goal became less important and security was the main guideline for the implementation of detention. More and more categories of inmates were excluded from rehabilitation activities (Boone, 2007).

Recently, the trend of impoverishment has been somewhat reversed with the implementation of the 'Prison Modernization' program in which rehabilitation programs, treatment, aftercare, training, individual approach and inmate labor have a central role. The evening program for inmates also made a re-entry. The treatment of inmates is sought in a combination of values such as support and respect on the one side, and values like structure and unambiguous rule enforcement on the other (based on Liebling & Arnold, 2004). The belief in the beneficial effect of humane treatment and rehabilitation activities seems to be back, although the drift of the argument is that the

expensive services and efforts of the prison should only be offered to motivated inmates who want to change their behavior.

The dynamics and experiences in the history of prisons, as well as political and public opinion, have led to changes in the prevailing view of human beings, criminals included. These changes each brought a different approach to crime which had an impact on the implementation of detention. Since the nineteenth century, more and more expectations, tasks and goals were attributed to prison sentences. Despite the fact that the goals of imprisonment will always be subject to political shifts of emphasis, the same goals have been more or less present for over a century in the Netherlands. The goal of rehabilitation has won and lost popularity in previous decades, but seems to be inextricably connected with the implementation of detention. The same goes for the goals of safety (of inmates, staff and society) and human dignity (amenities, privileges and treatment) that may be alternately brought to the foreground and background in prison policy. The themes of safety, human dignity and rehabilitation are also included in the mission of the Dutch prison system which reads: 'We ensure a safe and humane detention and work with our adjacent organizations and the inmate, towards reintegration. As a result, we contribute to a safe society.' (DJI, 2009)

Whether it is possible to categorize and concretize the purposes of detention and the tasks of Dutch prisons in laws and regulations (starting with the goals of the judge) will be analyzed in the next section. Wherever possible, this analysis is supported by empirical evidence from the literature.

2.3 Conceptual framework: from goals of imprisonment to implementation tasks

According to Kelk (2010) a judge seeks three main goals when imprisonment is imposed, namely: the restoration of the legal order (added suffering and recovery of the caused sorrow), individual prevention (aimed at the offender) and general prevention (aimed at society). These arise from combined penological theories in which both utilitarian and retributive perspectives have a place. This is called the 'unification theory' and has been the leading theory in the Netherlands as well as in many other

Western countries for decades (De Keijser, 2004; Hoekendijk & Kommer, 2011). In this theory, retribution is the reason for imprisonment (retributivism) while the time in prison should be used to rehabilitate the inmate (utilitarianism).

The three main goals include the following. By reciprocating the act of crime it is intended to restore the legal order (the offender is made to suffer by imprisonment). By doing so, the sense of justice of victims and society can be restored (Jonkers, 1975). Besides repentance of the offender and legal certainty of victims and society, another penal goal is at work here. The goal of *retribution* ascertains the legal certainty of suspects since the law sets a maximum on the sentencing per category of crime. Therefore, the judge has a guideline for the setting of criminal penalties in which the interests of multiple parties are guaranteed. When the boundaries of the penalty have been established – given the maximum sentence, the personality and the circumstances of the suspect – a weighing up of the two other main goals of imprisonment results in the judgment (Kelk, 2010).

Individual prevention focuses on the offender (and his or her offence) and has at least three functions, i.e. rendering someone inoffensive (one cannot repeat the offense: incapacitation), deterrence (one does not dare to repeat) and rehabilitation (one does not want to repeat). During confinement it is not possible for inmates to reoffend outside the prison. For the period of the sentence, society is safeguarded against a recurrence of the offence because inmates cannot leave the institution. Furthermore, the detention should contribute to the prevention or reduction of future reoffending because it is assumed that there is a deterrent effect in incarceration. The renouncement of reoffending is encouraged because the inmate is not willing to risk the same punishment again. Thirdly, a rehabilitative effect of detention is envisaged: the imprisoned person is prompted not to 'want' to reoffend.

With *general prevention* both deterrence and a moral standardization of society is pursued. The judge is sending a signal to society that a standard has been exceeded and that the excess leads to punishment. This punishment must impose a deterrent effect, so people are less inclined to consider a similar act of crime. Because the possibilities for escape are kept to a minimum, the deprivation of liberty has a general preventive effect as well.

The question that arises is: what can or should a penal institution concretely do with these somewhat abstract goals? Despite many studies on prison conditions and legal philosophical analyses regarding the importance and effect of the deprivation of liberty, little is published on this topic. Successively, we discuss what the three penalty goals mean for the implementation of prison sentences.

2.3.1 Main objective 1: Restoration of legal order

Does the retributive goal of a judge imply that the prison system must add extra hardships to its inmates? One of the few attempts to bring implementation tasks in connection to the goal of retaliation is undertaken by Lippke (2007). He argues that retaliation gives the most clues on how to implement a prison sentence. The extent (and way) of punishment, according to the retaliation theory, should be in proportion to the damage that the offender has done to the victim and society. According to Lippke, imprisonment is not justified in case of minor offenses. When deprivation of liberty is indeed appropriate, then the simple fact that one is incarcerated is sufficient to reach the goal of retaliation. The conditions of detention would therefore not have to add more deprivations than are strictly necessary for the proper detention of persons. Aiming at minimal restrictions for inmates is preferred because the suffering of the detention is in itself heavy enough. This is also reflected in the Dutch Prison Act and the mission of the Dutch prison system. These ideas encourage very restricted additional constraints by prioritizing human dignity. Indeed, the Dutch Prison Act states that inmates should not be subjected to any restrictions other than those necessary for the purpose of maintaining order and security in the facility (art 2, par 4), also called the 'principle of minimal restrictions'.

Besides retaliation, legal order may also be restored by giving inmates an active role in repairing the damage they inflicted on victims and society. Restorative detention can be implemented by making inmates aware of the impact of the offense. Also, they can be committed to restoring the damage to victims and society, such as raising money for a fund for victims (by inmate labor) and conducting so-called recovery meetings with victims (Blad, 2003). However, the Netherlands has a limited tradition with a restorative

interpretation of detention, although the subject seems to be gaining popularity in the current public debate and is also the subject of policy development.

2.3.2 Main objective 2: Can, dare and want not to reoffend: specific prevention

The penal goal of specific prevention is threefold. The first element, incapacitation (the 'can not' reoffend), needs no detailed explanation. Detention must be implemented so that inmates can not leave the prison (except for furloughs) during the time of their sentence. The security of society is ensured in such a way that inmates cannot commit an offense for the time they are in the prison. Obviously, inmates may commit offenses in the prison during their time within the walls and this is why the prison makes efforts concerning security and control. In terms of concrete implementation tasks this means that the fences and gates should be protected adequately and the import and export of goods and people are thoroughly controlled. These measures decrease the chances of escapes and violent disturbances within the walls.

The second element relates to the deterrence of offenders. Inmates should be deterred by the confinement and therefore refrain in the future from criminal behavior (the 'not dare' to reoffend). In line with the arguments concerning retaliation, the detention in itself is considered heavy enough and prisons should not add extra penalties. Actually, the deprivation of liberty means, among other things, that an inmate is: 1.) restricted in his or her freedom of movement, 2.) limited in the contacts with his or her social network outside the prison, and 3.) hampered by disruption of the continuity of employment or education. Apparently, no concrete and proactive prison tasks seem to arise from the deterrent purpose. The scarce empirical work on the topic suggests that 'forcing up' deterrent elements (in regime and security level) is counterproductive in terms of reoffending (Camp & Gaes, 2005; Gaes & Camp, 2009). Later on in this chapter we will discuss whether or not these issues are reflected in the current practice of the daily prison process.

The third element of individual prevention is the rehabilitation of inmates that should contribute to the 'do not want' to repeat offenses. The rehabilitation principle is established in the Dutch Prison Act and can be elaborated by an active 'approach' of inmates. Research has shown that the treatment style or approach of staff towards inmates can contribute to the realization of goals concerning prison conditions and

inmate behavior (e.g. Vuolo & Kruttschnitt, 2008; Reisig & Mesko, 2009; Liebling, Durie, Styles & Tait, 2005, Molleman & Leeuw, 2012). For instance, prison staff may encourage inmates to take responsibility and work on their rehabilitation. Also, prison time can be used for education, acquiring relevant skills and gaining work experience. All this may contribute to a successful re-entry into society.

In addition, it is assumed that the pursuit of minimal restrictions and limiting the negative side effects of detention can promote the rehabilitation of inmates and may reduce the chance of reoffending. In practice this is about mitigating or preventing the loss of income, jobs and housing, and disruption of social life. The principle of minimal restrictions (as addressed above) is enshrined in law and implies that the prison provides somewhat similar living conditions as those outside the walls. This can be expressed in amenities such as adequate care, leisure, sports, exercise of religion and freedom of information (the Dutch Prison Act describes minimal standards for these amenities). Because humane conditions are provided, detention causes less damage which may promote a crime-free future. With regard to much of the above amenities, research has indicated their contribution to a successful rehabilitation; however, a complete enumeration of the evidence would be beyond the scope of this chapter.

2.3.3 Main objective 3: General prevention

The third and final objective of a judge in imposing a prison sentence is general prevention. The question is how the prison system can contribute to sending 'a message to society that a legal standard has been exceeded'. The means to achieve the general objective of prevention is general deterrence (De Keijser, 2004). Because evading a prison sentence (the opportunity to escape) is restricted to a minimum, the deprivation of liberty has a general deterrent effect. The impossibility of escaping not only contributes to the achievement of the goal of incapacitation (as part of specific prevention), it gives a strong signal to society as well; in the case where a person is sentenced to prison, there is no way out. The resulting deterrent should dissuade citizens from committing crimes. The general preventive penal goal has no further 'operational' function in the daily practice of a prison. In the Netherlands, the principle

of minimal restrictions is in force as mentioned above. This means that (the length of) the incarceration should be the deterrent, not the conditions of the incarceration. A prison is not expected to actively send out signals to the wider community. The implementation of prison sentences does not focus primarily on society, but on the inmates: 'General preventive function is emphasized in the stage of threatening [by the legislature], the specific preventive function is emphasized in the stages of implementation (...) by officials of the executive authority.' (Pompe, 1950: 274) Later on in this chapter, we will discuss how to put these notions into practice and whether there is really no active signal given to the wider community.

Besides the daily supervision and treatment of inmates, the prison may inform society about what detention means, which may contribute to the goal of general prevention. Examples are the provision of information, guided tours (e.g. open-door days) and prison museums, enabling filmmakers and writers and the involvement of citizens in supervisory committees.

2.3.4 Prison staff

In addition to tasks that arise from the main goals of punishment, prison acts and regulations set specific requirements for the execution of imprisonment. Regulations in the daily operations that play a prominent role are embodied in the Occupational Health and Safety Act (OHSA) which states that every employer must create a safe workplace for its employees. Organizations, and prisons in particular, have therefore important tasks in minimizing safety risks and supporting their employees in their daily work. Prison staff must frequently deal with rebellious and obstructive behavior of inmates because the latter group is, among other reasons, limited in their self-determination and autonomy (Goffman, 1961). Moreover, Goffman argues that working with these people can be complex and burdensome because of the emotional engagement and empathy of the professional. The employer must make arrangements to enable staff to do their tasks in a safe manner. The safety can be enhanced by adequate staff training and regular prison cell searches for weapons and other illegal goods. The safety of staff is largely dependent on the extent to which he or she can count on colleagues when they are threatened or face risky situations.

The OHSA requires explicit attention to the psychosocial workload and absenteeism. Causes of stress among staff are workload, ambiguous tasks, uncertainty about the future and lack of support from the management (Kommer, 2009). A prison has therefore the explicit task to give attention to these aspects of working conditions. Dutch research has found empirical evidence on the connection between the aforementioned causes and the prevalence of absenteeism and the psychosocial workload (Bogaerts & Den Hartogh, 2008; Kunst, Schweizer, Bogaerts & Van der Knaap, 2008, Molleman, 2011a; Molleman, 2011b).

We restrict ourselves to the working conditions of the staff, as they play a crucial role in the process of the prison that revolves around working with people. Other themes also give rise to other tasks of the prisons, but are general features of operational management, such as the maintenance of the building (e.g. fire safety), a proper complaint system and financial health of the prison. The latter tasks apply to almost every organization and are not further specified here for the prison system.

The above-presented analysis can be summarized as follows in Table 1.

Objectives of criminal justice	**Goals of detention**	**Prison system tasks**	**Concrete subtasks**
restoration of legal order: - adding hardships / retaliation - recovery of harm to victims	- confinement - restoring sense of justice of victims and society	- see 'defusing of offenders' *(incapacitation)* - restorative detention	- making inmates aware of the impact of their crime, letting inmates work to pay compensation, bringing inmates in contact with victims (restorative conversations)
Individual prevention	- defusing of offenders *(incapacitation)* - deterrence - rehabilitation	-security and control - *unclear* - strive for minimal restrictions and prevent harm to inmates due to detention	- protection of fences and gates, surveillance of imports and exports of goods and people - preventing loss of income, jobs, housing and social contacts, doing training, acquiring relevant skills and gaining work experience, adequate care, leisure, sports, exercise of religion and freedom of information - encourage inmates to act responsibly in staff-inmate contacts and work on rehabilitation
general prevention	- setting standards - deterrence to society	- *unclear* - informing society	- e.g. open-door day for citizens, involving citizens with operation of the prison
	Other goals	- staff safety - acceptable psychosocial stress for staff	- staff training and searches for weapons, drugs etc. - collegial support - prevention of: high workload, job ambiguity, uncertainty about the future and lack of support

Table 1. From objectives of criminal justice to concrete tasks of prisons

A large number of tasks in the table are described in laws and regulations and are standardized, countable and sometimes expressed in terms of the availability and quality desired. However, this does not apply for some of the topics in Table 1 making it *unclear* how objectives of criminal justice can be translated into concrete implementation tasks. These topics include the principles of rehabilitation and minimal restrictions that raise dilemmas so that they may lead to differences in the quality of the detention between prisons and prison units. Furthermore, these objectives give rise to contradictory tasks.

2.4 Causes of the lack of clarity about prison tasks

In conclusion of this chapter, this paragraph addresses two causes of the lack of clarity about the tasks prisons have to perform.

In the Netherlands, imposing restrictions on inmates is only allowed if that is necessary for 'the purpose of the deprivation of liberty or in the interest of maintaining order and security in the prison' (article 2, paragraph 4 of the Dutch Prison Act). That is, only when the goals of imprisonment (which the law does not make explicit), or when order and security in the prison are called into question, restrictions may be imposed. The question then is why inmates are limited in the opportunity to control their own money, luxury food, privacy and freedom of movement within the prison and are held in simple prison cells. Such restrictions seem largely to have a different origin than the reasons which the law sees as a reason for restrictions. The currently imposed restrictions seem to be mostly related to retribution, deterrence, cost efficiency or to give a (political) signal to society (the detention should not be too luxurious). Van Veen (in: Boone, 2000) concludes that a prison sentence indeed should contain an element of suffering. According to Van Veen it follows from the Dutch Prison Act (art 2, par 2) that inmates should be rehabilitated; however, the character of the punishment should be maintained. The sparse case law does not expound how a prison should pursue such a 'character' (e.g. add some degree of retribution and deterrence). Between the turn of the century and 2012, the Dutch Council for Criminal Justice and Protection of Juveniles treated 16 unique complaints by inmates in which the Council used the argument of 'order and security of the facility' in its statement. However, the 'character' of the

punishment was not mentioned in the statements. Clarification on the precise operation of the principle of minimal restrictions is therefore lacking and can, as a consequence, cause differences of implementation between prisons.

Furthermore, the Dutch Prison Act stipulates that the prison sentence is implemented in such a way that it supports successful return to society of inmates as much as possible (art 2, par 2). Here, ambiguity exists in the phrase 'as much as possible'. The expression suggests that the possibilities are limited, but it is not mentioned by what (e.g. amount of money or effort, possibilities or skills of the inmates). Again, this may lead to implementation differences because local prison managers may interpret this section of the law in various ways.[6]

2.5 Conclusion

The themes of the mission of the prison system (safety, humanity and rehabilitation) are over 100 years old and do not seem to be greatly susceptible to changes, although the emphasis laid on the respective goals fluctuates in policy and implementation over time. The fluctuations are not only found over time; implementation differences are also found between prisons. At many points laws, rules and operating procedures give clarity about how Dutch prisons are expected to live up to the goals of the judge. On some points, however, this is not the case. For example, it is unclear how and to which extent prisons are expected to enforce rehabilitative activities. Therefore, it is problematic to measure the performance of Dutch prisons on these specific tasks.

[6] For some ideas on how the lack of clarity may be decreased we refer to the contribution in Delikt & Delinkwent by Molleman & Van den Hurk, 2012.

Chapter 3

A method to deal with dissimilar circumstances of public organizations in performance comparisons: evidence from Dutch prisons

This chapter previously appeared as: Molleman, T. & van der Heijden, P.G.M. (2013). A Method to Deal with Dissimilar Circumstances of Public Organizations in Performance Comparisons: Evidence from Dutch Prisons. *Public Administration Research*, 2(2), 1-14.

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 chapter, 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 measurements derived from the Dutch prison system. Performance analysts may use the plan to analyze performance of (semi-) public organizations.

3.1 Introduction

For several decades, companies and (semi-)public services have used performance measures to obtain information about the achievements of operational management (Radnor & Barnes, 2007). An additional aim for (semi-)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, 2010b). 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 (2012) 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, Romano, Tancredi & Klazinga, 2012; Silber, Rosenbaum & Ross, 1995; Landon et al., 1996; Staiger, Dimick, Baser, Fan & Birkmeyer, 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, Lovell, Schmidt & Yaisawarng (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 chapter 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:

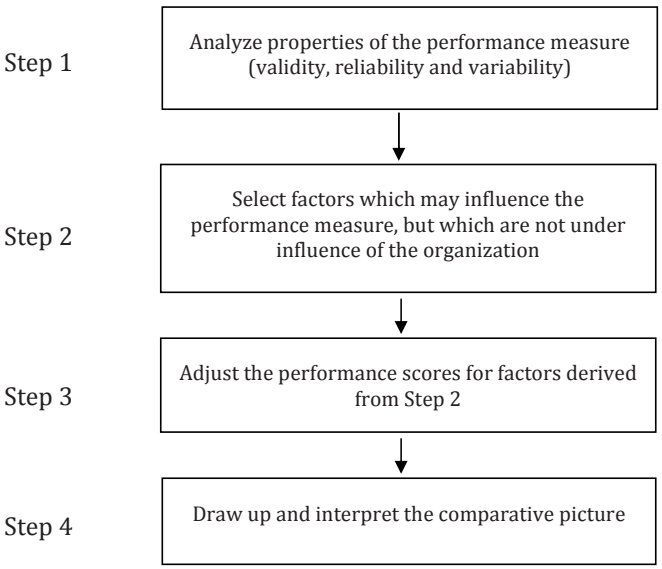

Figure 1. Step-by-step plan for performance comparisons

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.

3.2 Illustrative example: Dutch prison system

The Dutch prison system is a publicly run agency that operates by using an organizational performance monitor. In this monitor several measures are used to evaluate the goals of the agency 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 chapter, 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 the Appendix. 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, 2010b). However, this is beyond the scope of this chapter.

3.3 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.

3.4 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 chapter). 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, 2010b). 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.

3.4.1 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).

3.4.2 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, 2010b; 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.

3.4.3 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.

3.4.4 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(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 $\sigma^2_{prison} / [\sigma^2_{prison} + (\pi^2/3)]$. 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 local 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 $\sigma^2_{prison} / (\sigma^2_{prison} + \sigma^2_{unit} + \sigma^2_{ind})$.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.

3.5 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.

3.5.1 Illustrations for Step 2

For the selection of ND factors, we invited a delegation of prison managers (6 local prison managers and the head of agency) 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 local 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 local 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 local prison management cannot select his or her inmate supply. We aggregate the characteristics to the *prison level* because local 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.

3.6 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.

3.6.1 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 2). The regression model in Table 2 shows that most of the ND factors in the model reach statistical significance.

Violent incidents	B	S.E.	Sign
Level I variables (unit)			
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
Level II variables (prison)			
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

Table 2. 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.

Table 2 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).

We also run a multilevel regression model for the survey scale measure of *staff's feelings of safety*. Table 3 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.

Staff's feelings of safety		B	S.E.	Sign
Level II variables				
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
Level III variables				
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

Table 3. 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.

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.

3.7 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 et al., 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.

3.7.1 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.

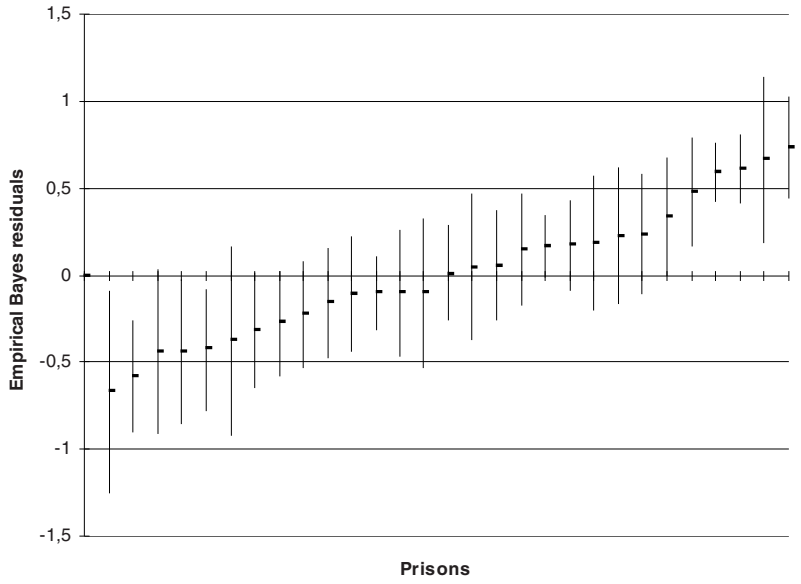

Figure 2. Caterpillar plot for the performance measures violent incidents. Confidence intervals set at 95%.

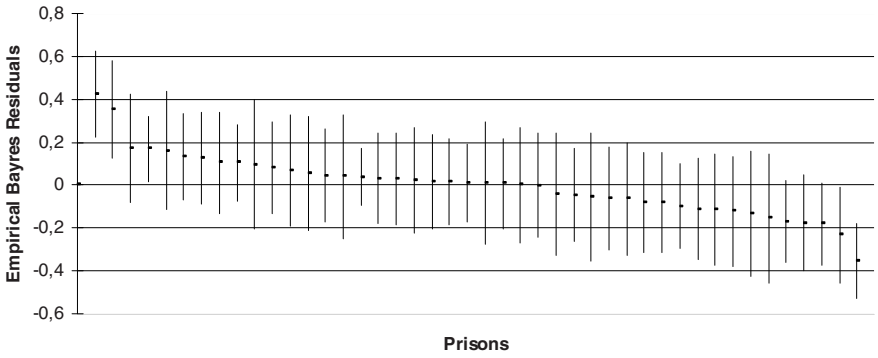

Figure 3. Caterpillar plot for the performance measures staff's feelings of safety. Confidence intervals set at 95%.

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).

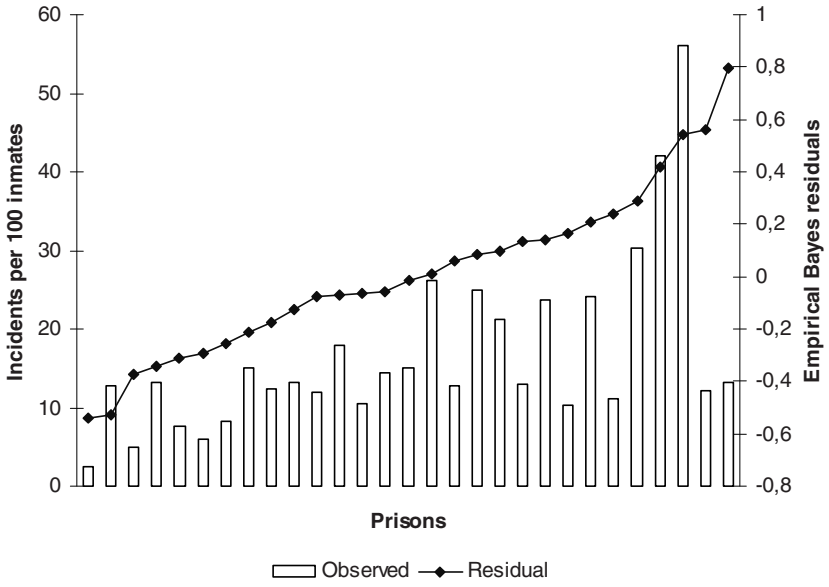

Figure 4. Observed scores and residuals for violent incidents. Prisons are ranked using their empirical Bayes residuals.

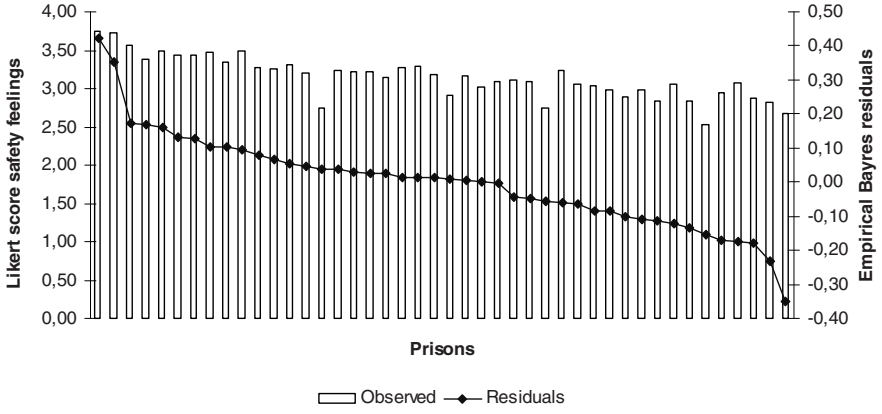

Figure 5. Observed scores and residuals for staff's feelings of safety. Prisons are ranked using their empirical Bayes residuals.

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 local prison management.

3.8 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 interpretation reasons 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, Gaes, Klein-Saffran, Daggett & Saylor, 2002; Molleman, 2008; Molleman & Leeuw, 2012). Survey data have been shown to be consistent with official prison records (Daggett& Camp, 2009; Molleman, 2011a). 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 skimming' 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, Rubenstein & Schwartz (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.

Chapter 4

Measuring performance in the public sector: towards a measurement strategy for composite indicators

This chapter is currently under review. Molleman, T., Leeuw, F.L. & Van der Heijden, P.G.M. (submitted). Measuring performance in the public sector: towards a measurement strategy for composite indicators.

Abstract

Performance measurement of (semi-)public organizations is not an easy exercise because the goals of (semi-)public sector organizations are often complex and stakeholders have diverse perceptions of what performance is. Therefore more and more often multiple data collection methods and composite indicators are employed to measure and evaluate organizational performance. However, a measurement strategy is currently lacking for using multiple data collection methods and composite indicators. A measurement strategy is provided based on critical realism. The safety of Dutch prison staff is used as an empirical example to test the convergence between measurements regarding outcome, mechanisms and context that are part of a theoretical network. In our example, measurements regarding outcome and mechanisms show links in the expected direction and are significant on more than one occasion. The theorized network was therefore partly confirmed but may be somewhat adjusted to measure prison staff safety more comprehensive in the future. The proposed measurement strategy showed to give guidance to measure (semi-)public sector performance and to create composite indicators that represent performance more comprehensively. It is argued that our ability to make valid and reliable assessments of (semi-)public sector performance has increased.

4.1 Introduction

Starting in the 1970s, New Public Management (NPM) became an important approach in public administration in the western world (Light, 2006; Hood, 1991). The purpose of NPM was to improve the efficiency and effectiveness of (semi-)public sector organizations, using private sector management techniques. 'Reinventing Government', authored by Osborne and Gaebler (1992), presented instruments for the development of entrepreneurial government, for example through the 'steering not rowing' principle. Other examples of entrepreneurial instruments include performance measurement, performance auditing, outsourcing, benchmarking and evaluations (Pollitt & Bouckaert, 2011). However, over the years many studies showed unintended side effects of NPM interventions (e.g. Van Thiel & Leeuw, 2002; Smith, 1995; De Wolf & Janssens, 2007; Pidd, 2005; Pollitt & Bouckaert, 2011; Leeuw, 2011; Pollitt, 2013). The side effects often have to do with mechanisms like 'perverse learning'. An example of perverse learning is the *tunnel vision* phenomenon in organizations: because attention is largely directed at the performance fields that are measured, other important issues pass into oblivion (Van Thiel & Leeuw, 2002). Another side effect of NPM interventions is *myopia*: because of a focus on direct results and outputs, organizations may forget longer term goals (Smith, 1995). Explanations for these side effects are attributed to 'implementation for the wrong reasons', such as to justify privatization (Gianakis, 2002) and a lack of 'champions'; managers and other employees have to spread 'the word' and stimulate actions at all levels of the organization to be more efficient and effective (Howell & Higgins, 1990; Kusek & Rist, 2004).

These kinds of explanations for disappointing NPM results all have to do with managerial and organizational behavior. However, and for this chapter crucial, there is another important group of explanatory factors, namely problems concerning the data collection methods (audits, archive, videotape and document analysis, surveys, etcetera) that are used for the measurement of organizational performance (Evans, 2004; De Wolf & Janssens, 2007). Since performance measurement may rely on different kinds of data that always contain – to some extent – a bias, their weaknesses may jeopardize the quality of performance assessment. Variability in performance between organizations

may therefore not solely result from the phenomenon under study (i.e. 'the' organizational performance), but from the way(s) of measuring as well.

For example, management reviews and performance audits can be used as data collection methods for performance measurement, but have the drawback that they may result in 'faking good' and 'teaching to the test': nothing seems wrong in the organization under review, because largely only successes or positive trends are shown while negative results are not reported or believed to be outliers (and hence are 'deleted'). While expectations seem to be met, in reality the performance can worsen. Other data collection methods, like client and staff surveys, are less likely to be confronted with these problems because the observer or reporter is not conscious of the goals of the data collection which creates less risk of *observer or subject bias* (Poister, 2010b). However, survey results may also be biased within groups: respondents may be strongly influenced by the opinions of others. Also, respondents may intend to exaggerate their answers in order to influence organizational management or for personal reasons (e.g. to 'cream' results). An additional problem with survey data is selection bias: a mismatch between the sample of the respondents and the relevant population yields a distorted picture. Instead of using surveys, it is possible to work with *agency records* from the organizations themselves. For the measurement of prison performance, for example, the presence of drugs in inmates is tested through urine checks.[7] However, what may happen is that the local prison management pushes inmates with a low criminal propensity profile forward for the tests, which means that the percentage of 'positive' cases will be understated. The selection of favorable 'inputs' is known as cherry picking. Related manifestations of this phenomenon are 'massaging the numbers' (Propper & Wilson, 2003) or 'creative accounting': managers may create arrangements (of money, means or results) that will be assessed more favorably by higher level management. To sum up, every data collection method has its strengths and weaknesses.

For the measurement of performance, all too often only one data collection method (or a *single measurement*) is used. Propper and Wilson (2003) give an example of such a single indicator approach referring to The New York Cardiac Surgery Reporting system.

7 This is seen as an indicator of prison safety because the presence of drugs may cause tension in the inmate population and deviant behavior of the drug user.

Hospitals and surgeons were ranked solely on inpatient mortality. By doing so, important performance issues such as hygiene or customer satisfaction were beyond the scope of the performance assessment. This may have had negative effects on the quality of care. Due to the aforementioned ranking system, surgeons and hospitals may refuse risky surgeries and may not give attention to good communication with patients. A Dutch example of using a single measurement concerns the – no longer applied – police targets regarding the number of speeding tickets and arrests of criminal suspects; every Dutch police officer had to reach a certain number of tickets issued in a year. This had as a negative side effect that a successful event without arrests would be interpreted as mediocre or even bad policing. In addition, the number of tickets does not address the quality and variety of police deployment (Vollaard, 2003) and more importantly, other police goals such as crime prevention may receive less attention because of the focus on only quantifiable aspects of police work. Other examples of single indicator comparisons in the areas of health and education are found in, for example, Normand, Glickman & Gatsonis (1997), Goldstein & Spiegelhalter (1996) and Woodhouse & Goldstein (1988). These examples show that there are considerable risks when single sources or measurements are used to assess the performance of (semi-)public organizations. As the tasks of civil services are (sometimes) multifaceted, single measurements are not appropriate (Selden & Sowa, 2004).

One of the solutions to this problem is to collect data by using a *multi-method approach*. Even though using several methods of data collection does not imply that measurement errors will be (completely) avoided, it is likely that weaknesses in certain data collection methods may be counterbalanced by strengths in others. Furthermore, the nature of the data (perceptual, behavioral, and so on) underlying the different data collection methods may have various foundations, such as different scientific philosophical approaches to reality and differences in validity and reliability.

These aspects need thorough consideration before we draw up a picture of organizational performance. A recent example of performance measurement using several data collection methods concerns a study that has been published on the construct validity of the concept of 'financial performance' in the business sector (Hamann, Schiemann, Bellora & Guenther, 2013). The authors defined and established

four different parts of financial performance, namely stock market exchange, growth, profitability and liquidity. Furthermore, they validated performance indicators and tested the interrelations between them on divergent, discriminative and nomological validity. Notwithstanding the accuracy of the methodology presented for financial performance, it is evident that organizational performance is more than only financial.

Performance of (semi-)public sector organizations often concerns a complex constellation of phenomena that can be perceived quite differently by various stakeholders. A methodology is therefore needed that is tailored to this specific character of (semi-)public sector performance. As an example, we focus in this chapter on the safety of staff in Dutch prisons.

The chapter is divided into four parts. First, we will adopt *critical realism* as the underlying epistemology from which relevant insights can be deduced to address measurement problems. Second, we consider how convergent and divergent results of performance measurements should be dealt with in a multi-method context. Third, which data collection methods may be used to assess (semi-)public sector performance and the performance of Dutch prisons in particular are explored. Fourth, we test the proposed method with empirical information obtained from the Dutch prison system.

4.2 Critical realism and organizational performance

Critical realism rests on the assumption that reality is stratified in terms of mechanisms, events and experience (Bhaskar, 1978; 2008).[8] Mechanisms are 'underlying entities, processes, or structures which operate in particular contexts to generate outcomes of interest' (Astbury & Leeuw, 2010, p. 368).

Since they refer to underlying or even hidden processes and structures, researchers *approximate* the mechanisms at work in reality with the use of theoretical

8 Critical realism is a general philosophy of science assuming the existence of an independent reality 'out there', but simultaneously questioning the way(s) in which we become knowledgeable about this reality (Pawson & Tilley, 1997; Maxwell & Mittapalli, 2010). Critical realists make an ontological distinction between scientific laws and patterns of events as the observer [and participants involved in a measurement] may influence the sequence of events (Bhaskar, 2008). Bhaskar intends the domain of the empirical (experiences) to be a subset of the domain of the actual (events), which in turn is a subset of the domain of the real (mechanisms).

constructs. A construct is a conceptual term used to describe a phenomenon of theoretical interest (Cronbach & Meehl, 1955). This phenomenon is built out of regularities that are interrelated and can be empirically tested (Carnap, 1956) by which partial 'regularities or constant conjunctions' of actual events may be assessed (Bechara & Van der Ven, 2011, p. 349). In short, links between observations of partial regularities (experiences) may identify events which may, in turn, empirically validate the existence of a (hypothesized) mechanism.

However, the data collection methods can be 'distorted epistemological lenses' that rely 'on flawed measures that yield imperfect empirical traces' (Edwards, 2003, p. 312; Cook & Campbell, 1979). Since observers (e.g. customers, managers, inspectors, registrars, staff and researchers) and their measuring abilities are among the causal factors of the sequence of events, we need to be aware of concomitant biases and artifacts. At the same time, we assume that different observations of these relevant actors may assess different (relevant) parts of a construct. A multi-method construct may include both quantitative and qualitative aspects of organizational performance and account for the perceptions of multiple relevant actors. This dovetails with critical realism since it allows for the consideration of a substantial range of sources to assess reality (Pawson, 2013).

Returning to performance measurement in (semi-)public organizations, constructs (in this respect called *performance dimension*) should include several data collection methods since in most cases different relevant stakeholders of (semi-)public performance have 'different perspectives on the world' (Maxwell & Mittapalli, 2010, p. 157). Politicians, policy makers and administrators sometimes tend to rely heavily on so-called 'hard data'. Performance analysts and evaluators, however, often recognize the advantages of simultaneous application of several data collection methods, mostly referred to as *mixed method research* (Johnson & Gray, 2010). Generally, multi-method studies are defined as studies that combine quantitative *and* qualitative methods (Tashakkori & Teddlie, 2010; Bryman, 2012). Although qualitative sources are informative, the daily routine of operational management and performance assessment will not allow for extensive qualitative measurement, reporting and interpretation. Especially if performance encompasses multiple facets and a large number of

organizations are included in a periodic analysis, quantitative information will be preferred. Moreover, quantitative measurements may assess qualitative parts of performance as well (i.e. surveys on quality).

In line with others (e.g. Kelly & Swindell, 2002; Selden & Soha, 2004) we argue that the advantages of a multi-method approach (quantitative *and* qualitative) can also be obtained with *multiple quantitative measurements* in cases where these are derived from *different data collection methods*. To sum up, critical realism and multi-method research are considered as suitable and compatible approaches for the comprehensive assessment of (semi-)public sector performance. The next question is: how can we use different kinds of data collection methods in such a way that we obtain a comprehensive picture of organizational performance?

4.3 Building composite indicators using triangulation

Our measurement theory implies that we have a valid and reliable view of performance if our data collection methods are capable of assessing a considerable part of the 'real' performance. Since the organizational mission statement (or goals) defines the themes by which an organization has to perform, a description should be developed to give guidance for a translation into performance dimensions with measurable elements. Therefore we map out the elements of a performance dimension and adopt a well-tried way to frame a phenomenon, which is called the nomological network approach (Cronbach & Meehl, 1955; Swanborn, 1973). The network is a framework which relates two types of 'nomologicals', namely theoretical dimensions and observable properties and quantities. Not all elements in the network need to be (directly) observable to test performance; the network should be drawn up in such a way that the measurements follow from the theoretical dimensions.[9]

We argued in the previous sections that the use of multiple data collection methods increases the likelihood that a performance dimension will be assessed in a

9 In the literature, terms such as theoretical constructs, measures, observables, perspectives, sources and nomologicals are sometimes used interchangeably in an unstructured manner. We take as a terminological basis that a performance theme may have more dimensions that can be made known with measurements from multiple data collection methods.

comprehensive way. Analyzing the similarity between the results of different data collection methods of the same object is called *methodological triangulation*,[10] mostly applied to double- or triple-check the picture that is given by results of a single measurement. In other words: 'If two tests are presumed to measure the same construct, a correlation between them is expected' (Cronbach & Meehl, 1955, p. 287), also called *consilience* (Wilson, 1999; Talbot, 2010). In case triangulation gives a *convergent* picture over multiple data collection methods, one might conclude that a certain part of the organizational performance in reality is indeed measured (Mathison, 1988).

But what if these measurements give a *divergent view* of performance in cases where these were expected to converge? Mathison (1988) pointed out the distinction between reliability and validity with regard to triangulation. On the one hand there is the *reliability issue*: performance information stemming from various data collection methods may diverge because bias in those measurements played a varied role. Therefore reliability of the data must be tested before further analysis. On the other hand there is the *validity issue*: various data collection methods may throw light on (slightly) different aspects of a performance theme.

A performance theme may be therefore subdivided in dimensions. Bechara and Van de Ven (2011, p. 350) pose the question 'how many dimensions are needed to represent the key features of a problem being investigated?' The authors suggest that the dimensionality of methods should match the dimensionality of the phenomenon observed, which is in line with the ideas of Campbell and Fiske (1959). The assessment of a performance theme therefore starts with the theoretical statement on dimensionality. Thereafter, whether the *classification* of the dimensions is reliable (and thus shows convergence) should be examined. Bechara and Van de Ven (2011, p. 350) also show that when different data collection methods 'converge on the same dimensions of a problem, this reliability provides confidence in having a valid representation of the problem domain'. In the cases where these methods lead to divergent views, this indicates that the dimensionality of a problem domain may not have been mapped in a valid and reliable way.

10 Sometimes other forms of triangulation are distinguished, like data and investigator triangulation. These are all subgroups of methodological triangulation since different data, theories and investigators can be involved in the methodological approach.

There is no commonly accepted guideline at hand on how to deal with the last-mentioned situation of divergence. In fact, Cronbach and Meehl suggested some sixty years ago that 'when observations will not fit into the [nomological] network as it stands, the scientist has a certain freedom in selecting where to modify the network. That is, there may be alternative constructs or ways of organizing the net which for the time being are equally defensible' (Cronbach & Meehl, 1955, p. 290). Moreover, the adjustment of a network may lead to a theory with a higher degree of corroboration (Popper, 1959).

Notwithstanding the logic of these ideas, it may remain tempting to erase or ignore outlying performance measurements in favor of 'polished' performance models. At this point, we may follow two options. First, the divergent measurement may be part of another dimension and there are arguments at hand to change the nomological net. The literature (*research repositories* like knowledge banks and systematic reviews) should be consulted in the search for alternative statements on the connections in the network. In case these repositories do not give a clue, we may consult experts or use our common sense. Second, the measurement may diverge because we discover reliability problems and other biases. Solutions then include the inspection of differences in data collection results within the *same* respondents (Dillman & Tarnai, 1988; De Leeuw, 2005) and data trials (Poister, 2010b) to check whether data are collected in the same way in every organization. Inadequate divergent measurements should be subsequently excluded from the network.[11]

Another issue in which we lack guidance is the degree of concordance that is desired between the measurements in a performance dimension. There are no general norms and thresholds for establishing a dimension and the calculation of composite indicators. Since we came to the conclusion that (semi-)public sector performance can be perceived differently by stakeholders, we do not expect beforehand high levels of convergence between measurements from different methods within one dimension. Confirmatory

11 In extreme cases it is conceivable that the apparently diverging measurement shows up to be trustworthy while the other (converging) measures point out to be biased (to the same extent or in the same way). Unquestionably, the decision on divergent measurementsshould not be based on data driven considerations.

factor analysis and the like are therefore not appropriate. Another approach is needed than that used for the development of performance dimensions with a relatively small scope and about which perceptions barely differ (e.g. financial performance, see Hamann et al., 2013).

We argue that measurements within a performance dimension should correlate with the expected sign and ideally in a significant way. In fact, the size of the correlation should *not be too high* because we want to prevent imbalance in the representation of the theoretical elements in a dimension (Hagedoorn & Cloodt, 2003). In case of high correlations, composite indicators may blow up small differences between organizations that 'may lead to an overemphasis on whatever it is that causes the two factors to change in similar ways' (Pidd, 2012, p. 239). Although controversial, we presume that adequate dimensions of (semi-)public sector performance have measurements that are correlated in the expected direction and do not exceed a 'medium effect size' of $r = 0.5$ (Cohen, 1988).

To build a nomological network for a performance dimension, we use a realist evaluation approach which suggests taking contexts, mechanisms and outcomes (CMO) into account (Pawson & Tilley, 1997). Outcomes (like consequences and results of organizational efforts) are central in the network; contextual elements (like the circumstances in which the organization has to perform) and mechanisms (like causal factors relevant for explaining performance) surrounding outcomes are used to validate the net. Moreover, measurements of mechanisms and contexts may be adequate indicators of (semi-)public organizational performance as well.

Using pairwise correlations we check the relations between the measurements. In the cases where the theorized links in the net are confirmed, a composite indicator can be calculated. In most cases it is necessary to scale the measurements before they can be added together.[12]

12 This can be done by standardization with z-scores (in case of normality) or non-parametric conversion techniques (in case of non-normality, e.g. Tehrani & Noubary, 2005). Next, the different performance measurements may be given weights to express their relative importance in a composite indicator. The relevance may be established by the trustworthiness of a measurement or by substantive considerations. Composite indicators are usually calculated with a simple linear equation (Pidd, 2012): $P = w_1x_1 + w_2x_2 + w_3x_3 + \ldots + w_nx_n$.

4.4 Sources of performance information

In this section we concentrate on data collection methods that can be used for the assessment of organizational performance and the performance of Dutch prisons in particular. We review some frequently employed methods in performance measurement and describe their qualities. We restrict ourselves to the often used data collection methods of *agency records, surveys* and *performance audits*. For the purpose of performance assessment we are in a fortunate position because all Dutch prisons: i) share one mission, ii) use centrally defined and established data collection instruments, and iii) have comparable organizational structures. This promotes possibilities for inter-organizational comparisons and straightforward interpretation.

When describing a source of performance information it is important to stress the role of the observer and/or registrar. Depending on these actors, an assessment can be made of the accuracy and integrity of the information. An observer may have a stake in how the scores are registered or may have limited facilities and possibilities to have knowledge of relevant information. In addition, the measuring of performance can be *obtrusive* or *unobtrusive,* which refers to the necessity of the observer intruding on the research context (Webb, Campbell Schwartz & Sechrest, 1966). This intrusion may disturb the natural sequence of events because his or her presence may influence people and make them act or answer in a socially desirable way. Nevertheless, these observations may be of importance due to their unique view on (parts of) performance in reality. Furthermore, measurements vary in their availability; some measurements are cyclically available (continuously, weekly, annually) and some data are irregularly collected. Some measurements are fueled by data collection methods following a strict framework with narrow definitions; others permit the interpretation of the observer. The measurement subject may refer to 'facts', acts, perceptions or attitudes. Moreover, information may throw light on quantitative or qualitative aspects of performance and may be measured at several levels of the organizations (e.g. individual, unit, organization).

---

where P is the score of the composite indicator, *w* is the given weight for performance measurement *x*.

Data collection method	Nature	Observer	Purpose	Unobtrusiveness	Availability	Qualitative / quantitative	Level
Agency records	Registrar, behavioral	Staff	Daily management	Medium / high	Daily	Quant	Organization / unit
Surveys	Self report, perceptual	Staff, inmates	Organizational improvement	Medium	Biennial	Quant	Unit
Performance audits	Check list, observational	Expert, inspector	Supervision, inspection	Weak	Irregular	Qual	Organization

Table 4. General features of data collection methods for performance measurement in the Dutch prison system

Table 4 presents some general features of data collection methods that hold performance information on Dutch prisons. Evidently, data collection methods in other organizations may have somewhat different features. We now describe the three main data collection methods of performance information and their ability to measure safety in Dutch prisons.

Agency records (also called administrative records or archival records) concern 'any data formally entered into an agency's record system by a representative of the organization' (Hatry, 2010, p. 243). Examples of these administrative records are financial information, number and type of complaints, realized production capacities and percentage of absent staff due to sick leave. A potentially substantial advantage is that these data are already available which eliminates the need for the collection of new data. Additional merits are the continuous availability and the fact that researchers do not need to intrude on the research context since the data is collected in the daily routine anyway (unobtrusiveness).

Among the demerits of agency records are the often encountered problems with missing data, accuracy of the data (e.g. through varying definitions or inconsequent

reporting), aggregation problems and unavailable data due to confidentiality and privacy (Hatry, 2010). Solutions to these problems are statistical imputation techniques, vocational training for registrars and data trials (Poister, 2010b). However, to a greater or lesser extent, bias in agency records cannot be avoided. For instance, violent incidents may be underreported because registrars may not register every incident (especially when management steers heavily towards low numbers of incidents) and recidivism rates are known for their considerable but unknown 'black number'. What is more, some records are subjected to intensive preparation of business controllers and other data editors in the organization. On the one hand this may contribute to accuracy, on the other hand this creates more opportunities for gaming or massaging the numbers (also see the 'mechanism of corruptibility of social indicators', in Campbell (1979)). A further disadvantage is that agency records are mostly quantitative and give limited information about the quality or situational elements of organizational performance. For performance measurement in Dutch prisons we use the following agency records from the registers of the year 2006-2007: registrations of violent incidents, urine checks on drugs, and staff's sick leave. Each prison's business controller made a judgment about the reliability of the registrations concerned. ("How do you assess the reliability of this registration?" Answering categories: very reliable, reliable, somewhat reliable and not reliable.) We only include organizations in our dataset in those cases where a controller assessed a measurement as at the least 'reliable'. Since these measurements have definitions agreed upon and have run for over a decade, we assume the included data are valid and reliable.

Surveys (also referred to as questionnaires) are used with increasing frequency for the assessment of organizational performance (Newcomer & Triplett, 2010). Customers, clients, patients, staff and other stakeholders of a (semi-)public organization may be asked to assess performance by answering surveys. Surveys can be performed with an interviewer by telephone or in a face-to-face situation. Furthermore, surveys can be filled in by respondents in a self-reporting situation using a computer – whether or not connected to the internet – or a paper and pencil version of the survey (Dilman, Smyth & Christian, 2009). All options have their pros and cons in terms of potential response, time and effort invested in survey development and analysis, and costs of the

measurement. Survey instruments are typically employed to evaluate perceptions and to determine priorities to improve organizational performance. The unobtrusiveness of surveys depends on the topic of the survey, but if the phrasing of the questions is defined free from value judgments and innuendo, respondents may not precisely know the reason for the research and are therefore likely to answer in a trustworthy way.

Surveys have the advantage that they are able to assess qualitative issues in a quantitative manner (e.g. questions on quality, answer expressed in numbers). They are often conducted periodically (e.g. biennially) and allow for analyzing trends in the survey results over the years. Talbot (2010, p. 41) argues that when the validity and reliability of the data is established, survey results are objective measurements of subjective perceptions of people. Moreover, a considerable response rate may yield a sharp picture of the organizational situation because a lot of observers are questioned about a certain topic of performance.

An important point is that respondents must have knowledge about the topics of a survey and have no stake in gaming their answers (to prevent *assessment* or *attribution* bias). In this study, we use the results of a staff survey of 2007 (N=1,689 correctional officers, response rate 63%) which is in line with the period the agency records refer to. The validity and reliability of the surveys, as well as the representativeness of the data, were tested exhaustively and led to satisfactory results (for a detailed discussion, see Molleman, 2011a). We adopt two staff survey measurements, namely scales concerning 'feelings of safety' and 'collegial support' (scale items in the Appendix).

Audit information concerns assessments of special mobilized external observers. They may be trained observers or experts working in audit teams or inspectorates. Ideally, independent auditors quantify 'conditions or behaviors that can be classified, counted, or rated by using one's eyes and sometimes other senses' (Berman, Brenman & Vasquez, 2010). For these assessments predetermined scorecards, checklists, or a framework of monitoring are used. Audits have a significant impact since they 'generate recommendations on which public officials can, and often do, take action' (Lonsdale, 2011, p. 14).

Despite these strengths, the supervision function of inspectorate bodies and audit commissions makes organizations conscious of being subject to performance judgments. Therefore the unobtrusiveness of the measurement method is weak. Another demerit of audits is their irregular availability; most likely, not every unit or organization of a (semi-)public service is audited annually. Audits can furthermore be costly to conduct because of the costs of audit personnel and the interruption of the daily production process through audit activities.

In this study we use the so-called External Security Audits (ESA) that are regularly held in Dutch prisons and which issue several elements of the security situation in prisons. We involved audit results that were conducted between March 2006 and May 2007 which ensures the data are related to the other data used in this study. Every audit element had multiple topics on which the auditors made their judgments. For each audit element we used the percentage of topics evaluated positively as a performance indicator (i.e. the percentage of the criteria that are met). We assume that the audit results are valid and reliable because experienced auditors with standardized checklists assessed the situation in the prisons. The elements of ESA that are included in this study are:

- Internal communication systems with the prison security center (10 topics): presence and functioning of the communication systems (for communication between staff and inmates in cell)
- Work instructions (13 topics): functioning of and familiarity with the working arrangement of tasks regarding security
- Incident logging (9 topics): notification of exceeding incidents
- Information security (22 topics): ICT, procedures concerning confidential information, e-mail and internet security
- Internal security (35 topics): checks of walls and fences, pre and post controls after using indoor and outdoor areas, daily cell inspection, checks of incoming and outgoing goods, inmate count and checks of locks
- Lighting (13 topics): inside and outside lighting, presence and functioning of regular and emergency lighting
- Contraband checks (20 topics): checks of incoming and outgoing inmates for control of contraband

In conclusion we may say that these data collection methods have diverse characteristics and therefore have the potential to comprehensively assess performance. We now take a last step to operationalize prison performance regarding staff safety and then look at the possibilities of establishing a performance dimension and an accompanying composite indicator.

4.5 Building the nomological network

The safety of staff in Dutch prisons is assessed by a network of elements derived from several data collection methods. The relations between the elements of the network are complex and subject to extensive scientific studies, often found in the penological and criminological literature. Although the relations between the elements are interesting in themselves, we focus here on those elements because they may together give a view of staff safety in Dutch prisons.

Outcome measurements are placed at the heart of the network. Closely related contextual elements and underlying mechanisms (Pawson & Tilley, 1997) are given an adjacent place in the network and may help to comprehensively assess the performance dimension. An exhaustive operationalization of the performance theme of prison safety is beyond the scope of this chapter and can be found elsewhere (e.g. Molleman & Van der Hurk, 2012; Wright, 2005; Logan, 1992).

The exercise of fitting measurements to theoretical concepts is generally arbitrary. However, as is argued above, a considerable part of reality may be assessed by including different data collection methods concerning outcomes, mechanisms and contexts.

In terms of *outcome*, staff safety can be assessed by their sick leave and the extent to which staff perceive the prison as a safe place. The latter measurement has the advantage that the staff survey is a direct way of questioning, but it has the drawback that it is open to biases such as socially desirable answering. A further outcome of staff safety is sick leave because job stress and aggression may lead to physical and mental problems among staff (see the systematic review by Schaufeli & Peeters, 2000). Such a

measurement is objectively obtained but has the disadvantage that sick leave has causes that are not job related, like influenza. These outcomes may result from certain *mechanisms*: underlying entities, processes or structures of daily management that contribute to staff safety. Collegial support, internal security, communication systems, information security, work instructions and incident logging are typical precautions that prisons take to ensure safety (Logan, 1992). When these processes and structures are substandard, the management is risking a high sick leave and unsafe feelings among staff while they have to do their job. With regard to performance measurement it is not without relevance that these factors are influenceable for a local prison management.

Finally, these mechanisms operate in particular *contexts*. Prison staff are expected to be less safe in a prison context in which: 1) contraband like weapons and drugs circulate, 2) inmates violate staff, 3) the lighting is not adequate and 4) many urine drug checks show positive. These context elements may even reinforce the mechanism elements in the net. This leads to the following nomological network (Figure 6).

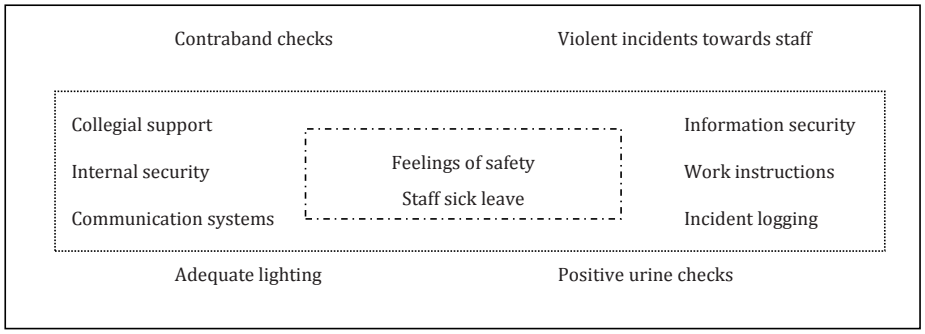

Figure 6. Theorized nomological network of prison staff safety

The outcome elements (feelings of safety and sick leave) can be found in the middle box; the mechanism elements are found in the surrounding box. The outer elements (contraband, violent incidents, lighting, and urine checks) form the contextual elements in the network. We do not attempt to develop an exhaustive model; there are conceivably more factors relevant to the outcome measurements. However, the presented network is thought to include the most relevant issues that can indicate the safety situation of prison staff. Furthermore, the elements can be measured with several

data collection methods with their respective strengths and weaknesses (see Table 1). We argue that it is not necessary to establish theoretical links (arrows) between the elements in the network. In cases where the elements are related in a logical way, we may have suitable ingredients for a composite indicator.

We use Kendall's Tau rank correlation procedures because we have a modest power in our analysis (44 prisons) and performance scores are sometimes equal for several prisons (Field, 2009).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
SS: Staff's feelings of safety (1)	1,00										
AR: Sick leave of staff (2)	0,01	1,00									
SS: Collegial support (3)	**0,21**	-0,14	1,00								
ESA: Internal security (4)	**0,29**	-0,15	0,09	1,00							
ESA: Communication systems (5)	**0,29**	-0,16	0,18	0,16	1,00						
ESA: Information Security (6)	**0,44**	-0,11	**0,33**	0,15	**0,23**	1,00					
ESA: Work instructions (7)	**0,26**	**-0,23**	0,12	**0,45**	**0,21**	0,13	1,00				
ESA: Incident logging (8)	**0,26**	0,03	-0,01	**0,41**	**0,22**	**0,32**	0,19	1,00			
ESA: Contraband checks (9)	**0,29**	0,09	0,18	0,12	**0,21**	0,17	0,12	0,13	1,00		
AR: Violence against staff (10)	0,03	0,09	-0,01	**0,22**	-0,06	-0,09	0,08	-0,06	**0,33**	1,00	
ESA: Lighting (11)	0,13	0,03	-0,01	-0,06	0,20	0,06	-0,07	-0,11	0,06	-0,01	1,00
AR: Positive urine checks (12)	-0,06	-0,02	-0,08	-0,05	-0,17	-0,15	0,07	**-0,28**	-0,03	-0,04	-0,03

Table 5. Kendall's Tau rank correlations concerning indicators of staff safety, bold coefficients meet p<0,05, N= 44 prisons

Table 5 shows that one of our central elements, the outcome measurement (1) on feelings of safety, has a significant relationship with all factors in the network that we have labeled as mechanisms. Moreover, the signs of these coefficients show significance and they are in the expected direction. Although the effect sizes are not strong, the table indicates that the outcome and mechanism measurements do not diverge. When we look at the context elements (9) to (12), it turns out that only adequate 'inmate inspection before/after transport' is related to staff's feelings of safety. The other effects are too small to reach significance and are not related to the outcome measurement. The outcome measurement of sick leave is only significantly related to the extent that work instructions are clear to staff and are complied with. However, the other (insignificant) relations with the mechanisms (3) to (6), are related in the expected direction (negative,

i.e. the better the precautions are implemented, the less sick leave). Similar to the first outcome measurement, the context elements are barely related.

To sum up, the outcome measurements and mechanisms show links in the expected direction and are significant on more than one occasion. Another important finding is that there are no opposing relations between the outcome measurements and the distinguished mechanisms. However, the current network has at least two weaknesses. Firstly, the outcome measurements 'feelings of safety' and 'sick leave' are not related to one another (r=0.01) which points to flaws in the theorized dimensionality. Sick leave seems to be quite different from feelings of safety on the work floor, possibly because sick leave largely has causes that are not job related. Secondly, the context measurements (10) to (12) do not play a meaningful role in this network. Apparently, adequate precautions (the mechanisms) are more strongly linked to staff safety than the concrete presence of dangerous and deviant behavior (contraband, drugs and violence).[13] Possibly these issues are seen as inherent to the job and problems only arise when staff are not adequately prepared and equipped for those risks. We argue that these context measurements may be relevant for other performance dimensions, for example *inmate safety*.

How to deal with the current network and should it be adjusted? We argue that the last three context measurements (10) to (12) have shown not to be part of the network. Because we established their reliability, this may indicate a validity problem. Arguably, we should seek for *other context elements* that are at work in reality. A look into scientific repositories shows that there is evidence that sick leave is independent of the culture of prisons (Lambert, Edwards Camp& Saylor, 2005). These authors showed that many of the variables found to be associated with sick leave in other branches also applied to correctional settings, such as job satisfaction, organizational commitment and job stress. Furthermore, staff sick leave has some connection with the here theorized mechanism elements (and in the expected direction), but it seems to vary independently from feelings of safety. In case we decide (on a matter of substance) to retain sick leave as a part of the network, we may give the measurement a lesser weight in the calculation

13 The factor 'lighting' has no relevant connection to any of the elements in the network. An explanation is that the measurement barely varies between prisons; all prisons meet over 70% of the requirements of this audit component.

of a composite indicator concerning staff safety (Eisenkopf, 2009). Because the mechanism variables consistently vary with the outcome measurements (and do not diverge), we argue that these are stable components to include in a composite indicator. Apparently, staff's feelings of safety appear as a central outcome measurement in the network because it has connections with mechanism and context elements stemming from other data collection methods. In this way, we may build a composite indicator in which several data collection methods are included and the contribution of the separate measurements is weighted on the basis of relevance to the nomological network.

4.6 Conclusions

This contribution focused on methodological opportunities for using multiple data collection methods and composite indicators to measure (semi-)public sector performance more comprehensively. Biases (like socially desirable answering and missing data) and unintended side effects (like tunnel vision and 'cherry picking') threaten the proper measurement of performance. We have seen that data collection methods such as agency records, surveys and audits have different strengths and weaknesses; the simultaneous use of these methods may reduce the risk of relying on a single (biased) measurement. Related measurements in a theoretical network contribute to the valid and comprehensive assessment of the often multifaceted (semi-)public sector performance. Next to the use of multiple methods concerning outcome, we argued to include in the network measurements of underlying mechanisms of the outcome and contexts following critical realism. With the empirical illustrations of one specific performance dimension, *staff safety in Dutch prisons*, we found evidence that elements of the theorized conceptual network converge. We did not find divergent measurements in the network; however, some measurements were not related at all. We have argued that in such situations the measurements may be part of another performance area (of prison safety) but are not directly relevant to the specific dimension under study. The network should be re-arranged and we may consider seeking alternative measurements to assess the dimension comprehensively. Since a bunch of measurements show convergence in our example, one can easily create a composite indicator in which the separate measurements may be given a different weight depending on their relevance to the network.

Did we make any progress in measuring (semi-)public sector performance? Probably we increased our ability to make valid and reliable assessments of (semi-)public sector performance in reality when we use measurements of outcomes, mechanisms and contexts. Moreover, the proposed methodology may ensure we assess a considerable part of performance which confirms our measurement theory. The established pattern of operationalizing and categorizing dimensions (however we only assessed one dimension here), theorizing a network, using multiple data collection methods and testing and adjusting the network, may contribute to the comprehensive assessments of performance. Convergent measurements give confidence that we uncovered some phenomenon in reality; however, we still cannot be entirely certain that the measurements are in accordance with reality. Therefore, we must realize that performance measurements are at best approximations and thus can be only indicative of organizational performance (therefore called *indicators*). Numbers do not tell the (complete) story, so explanations by managers and personnel themselves should accompany them. Furthermore, the use of composite indicators may help to provide user-friendly performance information and can be derived from empirically tested performance dimensions.

A lot of work remains to be done; for every multifaceted performance dimension in the (semi-)public service, such a deliberate analysis should be provided to assess performance comprehensively.

Chapter 5

The influence of prison staff on inmate conditions: a multilevel approach to staff and inmate surveys

This chapter previously appeared as: 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.

Abstract

The current study connects survey data of inmates and correctional staff in the Dutch prison system in order to describe and explain the impact of staff orientation and staff working conditions on perceived prison circumstances of inmates. Importation and deprivation theory are combined to test an integrated model to explain perceived prison conditions. By surveying staff (N = 1750) *and* inmates (N = 4673) independently within the same period of time and by afterwards *pairing* the results on the level of the housing unit (N = 173) using multilevel techniques, it is found that inmates´ perceptions of the prison conditions vary considerably between housing units. It is also found that staff's perceptions of prison conditions show congruency with those of inmates. Another important finding is that in housing units where the orientation of staff towards inmates is relatively supportive, inmates perceive their circumstances as more positive. Conclusions and directions for further research are provided.

5.1 Introduction

The objectives of a prison system vary over jurisdictions and may differ over time. The primary principle of 'keeping inmates inside' may be seen as a prominent function to contribute to the retaliation towards inmates and the societal repayment of debt. In addition, several other tasks and assignments are conceivable for a prison jurisdiction. A number of them are, to some extent, *almost always* present in corrections. Gaes et al. (2004) list the mission statements of the federal prison systems of the United States and Canada, and the missions of the prison agencies of the American Correctional Association. Most jurisdictions aim to make a contribution to the safety of society, inmates and/or prison staff, the reintegration of inmates within society, and the reduction of recidivism. Furthermore, twenty five state and federal American and Canadian jurisdictions (48%) pronounce that efforts should be made to treat prisoners humanely. In England and Wales, prisons aim to enhance the safety of inmates, treat them humanely, and reduce the risk of prisoners reoffending (www.hmprisonservice.co.uk). The goals of the Dutch prison system are similar to these; its purposes are to create prison conditions based on a dignified treatment of inmates, safety for society, staff, and inmates, and the reintegration of inmates (www.dji.nl).

The *existence* of mission statements does not *guarantee* the realization of these goals nor the prison conditions resulting from them. The implementation of a mission can be problematic and the execution of penalties is typically a 'work of man'. For those reasons, the *perceptions* of the 'prison atmosphere' of those who live and work in prison (inmates and prison staff) are important indicators of the prison conditions. Several determinants of these perceptions are considered in the literature. Not only the individual characteristics of an inmate may determine his or her perceptions of prison conditions, also environmental factors like the prison regime and the day-to-day treatment of prison staff may be relevant (e.g. Sykes, 1958). Several scholars refer to the importance of prison staff when achieving goals like safety, human treatment and re-integration of inmates (Arnold, Liebling & Tait, 2008; Bilby, 2008; Birgden, 2004; Bottoms, 1999; Hobbs & Dear, 2000; Liebling, 2000; Lindquist & Lindquist, 1997; Peterson-Badali & Koegl, 2002; Tewksbury & Mustaine, 2008). Garland (1990) even argues that 'operatives' of the penal system [e.g. prison staff], are the primary bearers of

the penal culture. Therefore, staff's orientation and working circumstances are relevant to the penal culture and the conditions of inmates inside prisons.

Research concerning the influences of prison staff on prison conditions typically focuses on staff's gender, age, and years in service or tenure (e.g. Cheeseman, Mullings & Marquart, 2001; Hemmens & Stohr, 2001), and concentrates less on factors such as staff orientation and staff's working conditions. Staff orientation refers to the attitude or treatment style of staff towards inmates; working conditions refers to factors such as contacts with the superior and the clearness of staff's tasks. There is a limited number of studies that looks into staff behavior or orientation when explaining inmates' perceptions of prison conditions. One study, based on a survey among inmates, focuses on the difficulty female prisoners have with adjustment to prison life. Inmates expressed to have more difficulty (e.g. experiencing a lack of privacy) when they receive no help from staff and feel that staff members do not go by the rule-book (Vuolo & Kruttschnitt, 2008). Another study, relying on interviews with inmates and official records, indicates that when inmates interpret the staff's authority as 'procedurally just', there are less inmate misconduct and rule violations (Reisig & Mesko, 2009). Liebling and Arnold (2004) reported that a respectful treatment by staff (as perceived by inmates) is highly correlated with various dimensions of prison life, such as perceptions of humanity. Similar results are found in relation to distress of inmates (Liebling, Durie, Stiles & Tait, 2005).

There is one other study, as far as we know, which uses staff *and* inmate data in order to explore what an effective staff orientation on inmate conditions could be like. In the 1980s, Nacci and Kane found that when prison staff is more satisfied, they also have a more helpful and supportive orientation, which is associated with lower records of sexual harassment among inmates (Nacci & Kane, 1984). What the mechanisms behind this finding are was not made clear.

Why is knowledge about the effects of staff orientation and their working circumstances on the prison conditions important? One reason is that it can be helpful in realizing a better understanding of the complex task of how to manage prisons in an effective way. Secondly, in order to do so, it is desirable that there is evidence about factors that, to some extent, are *manageable* or *malleable* by prisons and that contribute to the

realization of their goals is relevant. Therefore, the present study examines the impact of prison staff (their attitudes towards the orientation and their working conditions) on some important issues that matter to inmates, such as their perceptions of safety, humanity, and reintegration.

5.2 Explanations of perceived prison conditions

Some scholars are convinced that factors such as education, pre-prison experiences, genetics, personality, and cultural background determine current and future behavior of inmates (see Irwin & Cressey, 1962). Consequently, and according to *importation theory*, the influence of a prison on perceived prison conditions and inmate behavior is rather small, or even absent. Over the past decades, several studies have confirmed the assumption that characteristics of inmates (their criminal history, identification with criminal values, attitudes towards the legal system, the age of inmates, prior incarcerations, the length of the current sentence, their race, and their gender) are important predictors of inmate perceptions of prison conditions and deviant behavior (Bales, Bedard, Quinn, Ensley & Holley, 2005; Camp, Gaes, Langan & Saylor, 2003; Gendreau, Goggin & Law, 1997; Harer & Steffensmeier, 1996; Lindquist & Lindquist, 1997; Paterline & Petersen, 1999; Spivak & Sharp, 2008; Steiner, 2009; Steiner & Wooldredge, 2008; Wright, 1991).

Other scholars emphasize a direct link between the prison (as an institution) and the inmates' behavior (Clemmer 1940; Sykes, 1958). In *deprivation theory*, inmate's misbehavior (in prisons and after release) is seen as a response to the experienced prison conditions, including the deprivations that are inherent to confinement. Sykes (1958) lists five 'pains' of imprisonment, namely the lack of i) liberty, ii) goods and services, iii) heterosexual relationships, iv) autonomy, and v) security. Deprivation theory states that inmate subcultures (and misbehavior) are a reaction to the perceived pains of imprisonment partly due to the prison organization (e.g. regime and the treatment style of staff). Prison conditions can then be seen as a concept that is in-between characteristics of the inmates and their prison on the one side, and the behavior of inmates on the other side. Figure 1 may clarify these interrelationships. This chapter focuses on the links between these characteristics (both inmate and prison

factors) and the perceived prison conditions. It is very possible that prison conditions have an effect on post release inmate behavior, next to pre and post prison experiences of inmates. Nevertheless, the interrelationships with the current and future behavior of inmates (the most right box in figure 7) are no part of this study.

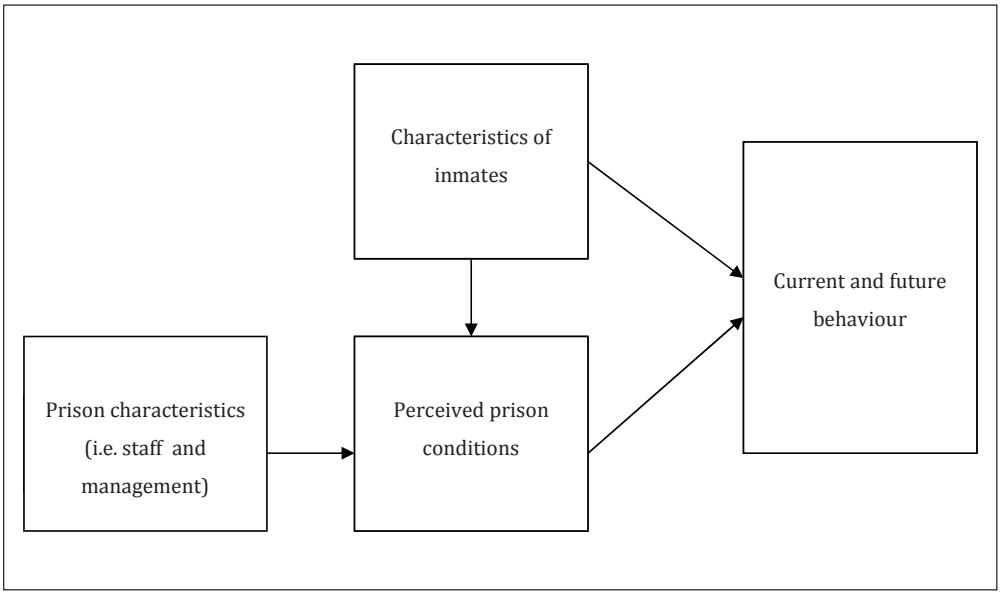

Figure 7. Hypothesized interrelationships between prison, inmate and staff characteristics, prison conditions and current and future behavior of inmates.

As a result, the prison organization can be seen as one of the *determinants* of perceived prison conditions, next to characteristics of inmates. Deprivation theory is sometimes viewed as the antithesis of the importation model, but also can be seen as a supplementary explanation which can be adopted in an integrated approach (e.g. Paterline & Petersen, 1999; Hochstedler & DeLisi, 2005; Cao, Zhao & VanDine, 1997; Huey Dye, 2010). Notwithstanding the relevance of importation factors, the chapter concentrates the attention on the influence of staff's orientation and staff's working conditions (as prison organization factors) on the perceived prison conditions of inmates.

5.3 The impact of the prison organization on prison conditions: staff as a determining factor

Different factors of the prison organization can be considered in an explanatory model of perceived prison conditions. In this section, we first discuss two general treatment orientation styles of prison staff and two leadership styles of prison superiors. Next, characteristics of a few demographics of staff and their working conditions are examined regarding their supposed effect on prison conditions. Finally, as an organizational predictor of prison conditions, the regime applicable to an inmate will be discussed.

Recently, an increasing number of scholars have stressed the importance of prison staff in explaining prison conditions, inmate well-being, and offenders' future behavior (e.g. Birgden, 2004; Liebling, 2000; McCorkle, Miethe & Drass, 1995, Tewksbury &Mustaine, 2008). Helpful and motivating staff (here also called: staff with a supportive orientation) are believed to contribute to the achievement of goals such as human dignity and reintegration. Staff who provides help when needed and creates a good atmosphere, is supposed to be an important factors in developing more positive perceptions of prison conditions, which are subsequently believed to discourage negative inmate behavior. Obviously, prison staff orientation is not a stable and static factor across prisons, housing units, and over time. Roth (1985) has demonstrated that institutional rules are enforced selectively, depending on factors such as inmate-staff relations and the mood a staff member is in. The orientation of prison staff may vary between a *supportive* style and a *rule enforcing* style.

Links between staff-inmate relationships and prison circumstances have until now rarely been examined. In an early study (DiIulio, 1987) it is shown that local prison managements differ in their assumptions regarding the use of official restraints to control and facilitate cooperation among inmates and prison staff. Therefore, DiIulio (1987) has defined three management concepts of prison superiors: the control model, the responsibility model, and the consensual model. The first model refers to working routines that are highly controlled by restrictions. The relations and communication between the superior and the staff, between staff members, and between the staff and inmates are rather detached. The second model refers to informal interaction and is not

based on restraints, counterforce, and sanctions. In the responsibility model, staff prefer informal contact with inmates above rule enforcement. The third management approach is about a mixture of the former two models. The former two models of prison management fit in with transactional and transformational leadership theories. A transactional leader builds his contacts with his staff on a social exchange system of rewards and punishments (Burns, 1978).

The primary focus of this type of leadership is the accomplishment of staff's short run tasks (here called *directive leadership*). Transformational leadership focuses on *faith* rather than tasks and economic interests (Northouse, 1997; Bass & Steidlmeier, 1999). A transformational leader tries to create a motivating relationship with his or her staff by assigning responsibilities and empowerment (Aronson, 2001), in the present chapter called *entrusting leadership*. Camp et al. (2003) have argued that the managerial approach of superiors can have an impact on the institutional culture of a prison. There is some evidence suggesting that a 'command and control' leader may have a detrimental effect on the achievement of several goals of prison, especially rehabilitation (Craig, 2004). The explanation is that when a supervisor follows the rules and gives strict penalties, inmates will lack the possibilities to practice and function in a 'normalized situation'. Consequently, the relationship between staff and inmates leaves little room for discussion and practice: even a small *faux-pas* of an inmate may lead to sanctions and will give inmates fewer opportunities to reintegrate. However, supervisors who manage on the basis of trust and social discretion are expected to contribute more to realizing prison goals.

Empirical research indicates that the different management approaches distinguished by DiIulio result in different working conditions for staff members. The management style of superiors is found to be a good predictor of job satisfaction and role strain among correctional staff (Reisig & Lovrich, 1998; Stohr, Lovrich, Menke & Zupan, 1994). However, not only staff may perceive the effects of the management style of their superior, inmates may *also* notice consequences. There exists little empirical evidence on the direct link between management approaches and inmates' behavior. Two studies suggest that discord between staff and superiors is connected to deviant behavior of inmates, namely collective disorders and homicides (Reisig; 2002; Useem

and Reisig, 1999). Except for such events, it is unknown how different managerial styles may influence inmates' perceptions of prison circumstances.

Also human capital factors may be important here, since we expect that the appreciation by staff members of their job will affect their job attitudes and will also have repercussions on prison conditions. For instance, a prominent issue in the literature includes the ambiguous tasks prison officers have to perform and the risks and difficulties involved. Staff members are expected to ensure security and control. *At the same time*, however, they are supposed to help inmates with their social reintegration process. Conflicting roles can lead to stress and disillusion about the possibilities of reintegration (Craig, 2004; Farkas, 1999; Liebling, 2000; Reisig & Lovrich, 1998; Tewksbury & Mustaine, 2008). Therefore, we argue that human capital issues are positively related to the ability of staff to implement activities focused on safety, humanity, and reintegration.

Next to the staff and management factors mentioned before, demographic characteristics of prison staff need to be addressed. These characteristics are found to be related to staff members' support for the realization of detention goals, such as rehabilitation, retribution, incapacitation, and deterrence (Tewksbury & Mustaine, 2008). In line with several empirical studies (Camp et al., 2003; Farkas, 1999; Reisig & Lovrich, 1998; Steiner, 2009; Tewksbury & Mustaine, 2008), the following staff characteristics are especially relevant: education, work experience, and gender.

Perceived prison conditions may be subject to differences stemming from variations in prison regime. For instance, remand prisons differ from prisons in the availability of amenities, like labor and social programs. Inmates in remand prisons are waiting for trial or transport, and are not extensively provided with social programs and special care, unlike prison inmates (Lindquist & Lindquist, 1997). These differences may also have consequences for the correctional orientation of staff since the goals of regimes have various characteristics. Therefore, the correctional orientation of staff is also expected to vary over regimes (see also McCorkle et al., 1995; Wright, 1991). In the Netherlands, the regime an inmate is attached to can differ within a *correctional facility*, and on occasion even *within a housing unit*.

5.4 Method

To assess inmate conditions and staff factors, one can use different methodologies and types of data collection, such as official records and observations. We argue that survey measures among inmates are suitable to evaluate prison conditions. The fact is, inmates are continuously exposed to those conditions. One might suppose that inmates report more negative perceptions about their circumstances than reality would allow. However, we do not have arguments why this *tendency* would differ substantially between individual prisoners. The way of answering of inmates to survey questions varies even in 'a systematic fashion' between *prisons* (Camp, 1999), which is a strong indication that this data source reflects the prison conditions of a particular facility.[14]

Staff factors are obtained with a survey among correctional officers. Bringing these inmate and staff measures *together* can be done by creating a *second level* on the inmate data with averaged staff survey measures, which is illustrated in the subsequent section.

5.4.1 Instruments and variables: Inmates

Sykes' pains of imprisonment, as well as the general prison goals of safety, humanity and reintegration, form the essence of the concept of prison conditions. Deprivations concerning liberty and security may be reduced by guaranteeing *autonomy*, *clarity of the prison rules*, and the *safety* of inmates (the goals of safety and humanity). By making efforts on the *availability of activities* during the day, *program delivery*, and possibilities for *contact with the outside world*, prisons may contribute to reducing the 'pains' concerning personal relationships and goods and services (the goals of humanity and reintegration). The use of an inmate survey based on the Prison Environment Inventory (Wright, 1985) allows us to obtain data on these topics. The questionnaire's scales are adapted and applied to the Dutch situation; they have been tested extensively on their

14It is possible that inmate *perceptions* collected in a survey do not exactly reflect the *actual situation* in a prison facility. Results that are based on self-report data obtained from offenders can be susceptible to deception. It is argued that offenders possibly lie, fake assumptions, and cannot be trusted (Camp, 1999; Mills, Loza & Kroner, 2003; Kroner & Loza, 2001). However, inmates agree on their assessment of the prison conditions since they answer in a systematic fashion (Camp, 1999). Paulhus (1984) distinguishes two 'natures' of social desirability in answering (survey) questions, namely self-deception and impression management. As we investigate *perceived conditions*, we do not expect *favourable* representations by inmates.

psychometric properties (DJI, 2004; Molleman, 2008). The Dutch version of the Prison Environment Inventory contains twelve scales, all consisting of items with a five-point Likert scale, varying from strongly agree (1) to strongly disagree (5). In this study, six scales are used that are discussed below. The reliability of the scales is good (Cronbach's alphas are all above .73).

The measures of safety are represented in a scale containing items about feeling safe in general in the prison institution, having been threatened by fellow inmates or staff, and having been exploited by other inmates. Autonomy includes measures on the degree to which inmates think that they control their own affairs, whether staff stimulate this, whether staff are receptive to their opinion, whether inmates think their lives are completely restricted, and whether they are treated like adults. The monotony scale measures the quality of daytime occupations, of weekend-, evening-, and leisure time, of the time spent in cell, and the issue whether an inmate can amuse himself in cell. Program delivery is measured by asking inmates to reflect on their satisfaction with the available labor, education, sports, library and recreation. The scale concerning contact with the outside world encompasses issues such as the possibilities to keep in contact with one's family, friends, and lawyer, the facilities for receiving visitors and attending court to witness the progress of one's case, and the guarantees for privacy during visits and phone calls. The scale clarity of rules and rights has statements on the clarity, communication, and sanctions concerning the prison rules, and the clarity and communication regarding the inmates' rights. These six scales represent the perceived prison conditions and make up the dependent variables in this study.

Personal characteristics of inmates are age, gender, prior incarcerations, length of the current sentence and cultural background.15 Finally we have included, a measure for regime. Three regimes are distinguished, namely prisons, remand prisons and an open regime. In the Netherlands, different regimes are found within the same facility. Except for those differences within facilities, Dutch prisons and remand prisons both generally have the same security level and capacity, and provide basic services like a library, religion and recreation. For those reasons, the Dutch remand prisons may be not

15We used the variable *cultural background* since *country of birth* is not sensitive to second- and third-generation migrants. *Cultural background*: We put together missing background and other cultural background, since we assumed that respondents with other or mixed cultural backgrounds have difficulties in answering this question.

exactly comparable to jails or remand facilities in other countries like the United States. In the Netherlands, the prison system distinguishes, besides prisons and remand prisons, an 'open' regime that supplies a relatively large amount of reintegration efforts. These facilities without bars, particularly intended for inmates at the end of their sentence, have a low security level. It is expected that inmates in 'prison' and 'open' regimes perceive their conditions more positively compared to inmates in remand housing units.

5.4.2 Instruments and variables: Staff

The measures of staff are derived from the 21 BASAM staff inventory scales (BASAM: Basic Questionnaire Amsterdam, see Biessen & De Gilder, 1993) and nine additional prison scales, all with satisfying psychometric properties (Cronbach's alphas are all above .70, see further Biessen, 1992; Molleman, 2008). The questionnaire used five-point Likert scales, varying from strongly agree (1) to strongly disagree (5). The scales are compiled by averaging the clustered propositions. For our analysis, we were especially interested in seven scales of the staff survey. First, the scale is used that measures staff's *supportive orientation*. This scale addresses items on the degree to which the staff i) do not begrudge the inmates, ii) prefer helping inmates rather than rule enforcement, iii) are willing to give individual help, and iv) involve inmates in matters they are concerned about. Second, the scale *rule orientation* is adopted, which includes propositions concerning the extent to which staff generate clarity about the prison rules, staff think it is obvious what inmates can expect from them and vice versa, and whether it is clear to inmates when they violate a rule. Third, measures of *meaningfulness of the job* is derived from items regarding staff's perception on whether they find their work valuable and whether they feel proud to be doing their job. Fourth, a scale called *perceived responsibility* is adopted. This provides an insight into the degree of responsibility that the staff experiences, into whether they consider the results they achieve in their work to be caused by their own efforts, and into whether they consider to be doing their job well. The degree to which staff say that they want to be involved in solving problems that occur during their shift is also included in this scale. Fifth, the scale *conflict of function roles* measures whether staff know what their function requires from them, whether they know what their colleagues expect from them, whether it is

clear what they are supposed to do during working hours, whether they know how to achieve results, and whether they know what they are responsible for. Sixth, the scale *directive leadership*[16] refers to prison managers who check that the work is done on time, criticize poor work, and reprimand staff when they perform beneath acceptable norms (or inferior norms in comparison to other staff). The seventh scale which is included into the analysis, *entrusting leadership*, contains questions about whether a local prison manager delegates responsibilities and trusts his staff to act appropriately. In this management style, rules, procedures and routines are less necessary. The manager consults his staff about the allocation of tasks and lets staff have a say in issues that concern staff members, for instance regarding the working schedule. Finally, characteristics of prison staff are measured; these include their *education*, *work experience* and *gender.*

All staff variables were aggregated to the housing unit level by means of averaging. Univariate specifics of the operated variables are found in table 1. All variables are tested for collinearity and multi-collinearity. Correlations between all independent variables show to be far below the r=.65 cut off and auxiliary regressions show lower R^2-scores than for the original regression, using Klein's rule of thumb. Moreover, additional 1/VIF-tests do not indicate multi-collinearity, since results do not show low proportions (all above 0,40).

16 The scales directive leadership and entrusting leadership are measures of the attitude of a staff member towards his or her direct supervisor.

	N	Mean	*SD*
Level I (dependent variables)			
Safety	4309	3,45	0,67
Autonomy	4372	2,96	0,76
Monotony	4411	2,51	0,84
Program delivery	4270	3,04	0,91
Contact with the outside world	3766	3,14	0,91
Rights and rules	4302	3,03	0,84
(independent variables: inmates)			
Age	4437		
≤ 19 years	181 (4%)		
20-24	758 (17%)		
25-29	752 (17%)		
30-34	723 (16%)		
35-39	734 (16%)		
40-44	534 (12%)		
45-49	334 (8%)		
50-59	334 (8%)		
≥ 60	87 (2%)		
Prior incarceration (0=no, 1=yes)	4000	0,56	0,5
Total sentence	4153		
< 3 months	647 (16%)		
3-6 months	837 (20%)		
6 months - 1 year	770 (19%)		
1-3 years	683 (17%)		
3-5 years	775 (19%)		
> 5 years	441 (11%)		
Cultural background* (Ref.: *Dutch*)	2295 (49%)		
Turkish	243 (5%)		
Moroccan	250 (5%)		
Surinamese	331 (7%)		
Antillean	254 (5%)		
Other	1300 (28%)		
Gender (0=man, 1=woman)*	4673	0,08	0,27
Regime (Ref.: R*emand prison*)	3094 (66%)		
Prison	1129 (26%)		
'Open'	260 (6%)		
Level II (independent variables: staff)			
Education*	173	4,51	0,73
Work experience	173	13,57	5,02
Gender (0=man, 1=woman)	173	0,23	0,19
Supportive orientation	173	3,4	0,34
Rule orientation	173	3,76	0,22
Meaningfulness of the job	173	3,64	0,35
Perceived responsibility	173	3,99	0,23
Conflict of function roles	173	4,02	0,25
Directive leadership	173	3,04	0,44
Entrusting leadership	173	3,14	0,48

Table 6. Descriptive statistics[17]

17* The measure of staff education as part of the staff survey is categorized as follows: 1 = primary school ascending up to an academic degree (=8). The variable is used at the aggregate level of the prison unit.

5.4.3 Data collection and criteria for inclusion

In the spring of 2007, within a period of two months, the Dutch prison system conducted surveys among inmates and staff *in all 48 facilities* (with a mean capacity of 347 inmates per facility). Every inmate and staff member had the opportunity to fill in the questionnaire.[18] In this study, only prison workers are included who have direct contact with the inmates on housing units. Both surveys had an overall response slightly below 50%. Respondents without a housing unit code are excluded from further analyses. Since staff data are aggregated to the housing unit level, we also control for outlying effects by erasing all units with less than five responding staff workers.[19] A number of 57 housing units and 5 facilities fall out of the analysis due to this criterion. Having applied these criteria, 1750 prison workers are coupled (41,8% of the population) to 4673 inmates (37,9% of the population). The included prisons varied from two to twenty housing units after applying the criteria. A housing unit has on average 10,1 staff members and 27,0 inmates. The respondents, staff and inmates, are nested in 173 housing units, which fall under 43 prisons.

Tests are carried out concerning the representativeness of the respondents before and after the removal of excluded respondents. Responding inmates are examined on their age, ethnicity, gender and length of sentence. Staff workers are examined on their age, gender, tenure, and amount of contract hours. Before and after the application of the criteria for inclusion, the respondents represent the inmate and staff populations (for extensive reporting see: Molleman, 2008).

5.4.4 Pairing Survey Results and Hierarchical Models

A growing number of empirical prison studies use *multiple data sources*, aimed at increasing the explanatory power of a study. In prior research, registered data have been combined with inmate population records (e.g. Bales, Bedard, Quinn, Ensley &

Therefore, individual information on the distribution of the categories cannot be given. In the original individual level staff dataset (N=1750), the modal category is intermediate vocational education

18 Participation in the surveys took place on the basis of voluntariness and anonymity. Questionnaires for inmates were distributed in each prison cell before locking the doors at night and collected the day after. No rewards were given for filling them out. The inmate survey was available in eight languages. Staff questionnaires were filled out and collected during team meetings, in absence of the staff's superior. Staff who were pregnant or ill were sent a questionnaire to their homes.

19Units with less than 5 responding inmates were also excluded.

Holley, 2005; Camp et al., 2003). Furthermore, inmate population records have been related to survey data (e.g. Wright, 1991). In the present study, inmate and staff survey data are *coupled* to integrate two independent perceptions on the same workplace and within the same housing unit. In order to couple the surveys, each respondent, inmate and staff, has a unique code for the housing unit and the prison they live or work in[20]. All staff measures are averaged on the housing unit level, and attached to the individual inmates. As a result, each single inmate within the same housing unit is assigned to the same (averaged) staff variables.

The questions whether and to what extent both staff and inmate perceptions are congruent with regard to the conditions inside a prison has only very seldom examined. As far as we know, only Camp et al. (2002) found that perceptions of prison staff and inmates on the sanitation in the dining hall and the housing unit are highly congruent. In order to examine whether connecting of the two surveys can also yield meaningful knowledge here, we test the analogy between a couple of staff and inmate scales dealing with the same subject (perceptions of *contacts between staff and inmates* and perceived *hygiene*). The scales are correlated on the housing unit level (N=173 housing units). First, inmates are asked about their perceptions of their contacts with staff. Staff are questioned on how they think about their contacts with inmates. It appears that both perceptions are significantly and positively correlated ($r=.36$, $p<0.01$). Furthermore, the relationship between inmates' appreciation of the hygiene in the prison facility and staff's perception of the physical working conditions (i.e. items on cleanliness, moistness and temperature) is tested. As expected, both perceptions are significantly correlated ($r=.20$, $p<0.01$)[21]. The differences in the legal and social status of staff and inmates do not seem to lead to rival responses.

By adding *averaged* staff variables to the individual inmate data, we have created a *second level* in the dataset and are able to account for influences of the *prison housing unit*. Although there is no information available on variables at the *prison* level, housing units are nested into a prison. Since each level is a potential source of variability (Hox,

20 Unlike prisons in some other countries, prison staff in the Netherlands is consistently placed inside housing units and does not substantially rotate. Therefore, the housing unit is an appropriate level of analysis.

21 The coefficients are not notably substantial. Plausibly, this is partly due to the slightly different questions in the survey instruments on the topics.

2010; Snijders & Bosker, 2012), the relative variance of the prison level is examined but does not exceed an intra class correlation (ICC) of 3% in the empty models. Therefore, the recognition of two levels (individual inmate level and the housing unit level) and the use of hierarchical level models is appropriate. All multilevel models in this study are tested against ordinary least squares models and showed up to give a significant improvement. The procedures of hierarchical models represent an approach taking into account the social contexts, here referred to as staff and management characteristics, as well as the individual actors(Snijders & Bosker, 2012), like inmates' characteristics. The analyses used in the present chapter allowed for random intercepts for each group and, as we expected them to differ, fixed slopes are applied since general relationships are hypothesized. The statistical analyses of the data were performed on Stata 11, using multilevel mixed effects linear regression models. Besides the coefficients, ICC rates are reported for each of the models. The ICC expresses the degree of resemblance between micro-units belonging to the same macro-unit (Snijders & Bosker, 2012), and is used as an indicator of the amount of the total remaining variance attributed to the distinguished levels.

5.5 Results

Table 7 contains the results of the multilevel analyses concerning perceived prison conditions. On level 1, characteristics of inmates show significant coefficients. Older inmates are more satisfied with their daytime occupations (*monotony*),their *contacts with the outside world*, and the *clarity of rights and rules*. Conceivably, older inmates accept their fate more easily. Those who have been previously incarcerated are less likely to appreciate the *program delivery*, but they value the *clarity of the rules and rights* as more positive. These findings are also in line with the expectations since re-incarcerated inmates are already familiar with the prison rules and the program components may often have been repeated. This last argument is also applicable to the finding that the length of an inmate's sentence is negatively connected to perceptions of program delivery. Furthermore, some ethnic minorities feel less *safe* compared to autochthonous inmates. Inmates falling into the category 'other cultural background' perceive the *program delivery* and the *clarity of the rights and rules* as relatively

negative. Possibly, problems with communication, customs and language explain those differences. The significant links between the *regime* and perceived prison conditions are as expected, as prisons and open facilities permit more time to be spent outside the cell and facilitate more activities, such as labor.

	Safety	Autonomy	Monotony	Program delivery	Contact with outside world	Rules and rights
Level I (independent variables: inmates)						
Age	-0.05	0.02	**0.13***	-0.02	**0.09***	**0.08***
Age ^2	0.01	0.00	-0.01	0.00	0.00	-0.01
Prior incarceration (1=no, 2=yes)	-0.01	0.03	0.05	**-0.07***	0.02	**0.09***
Total sentence	-0.01	-0.01	0.01	**-0.04***	-0.01	-0.01
Cultural background (Ref.: Dutch)						
Turkish (1=yes)	**-0.22***	-0.07	-0.06	0.00	0.02	0.09
Moroccan (1=yes)	-0.08	-0.13	-0.10	-0.12	0.08	-0.06
Surinamese (1=yes)	**-0.14***	-0.10	-0.03	-0.06	-0.04	-0.10
Antillean (1=yes)	-0.03	-0.05	0.11	0.06	0.04	0.01
Other/mixed background (1=yes)	-0.07	-0.08	0.02	**-0.13***	0.00	**-0.13***
Gender (1=man, 2=woman)	-0.09	0.06	0.01	0.06	-0.03	-0.02
Regime (Ref.: R*emand prison*)						
Prison	-0.06	0.09	**0.29***	**0.16***	0.12	-0.02
'Open'	-0.06	0.08	**0.48***	0.09	-0.16	0.03
Level II (independent variables: staff)						
Education	-0.01	-0.03	-0.02	-0.03	-0.01	0.02
Work experience	0.00	0.00	0.00	0.00	0.01	0.00
Gender (1=man, 2=woman)	-0.05	0.00	0.05	0.07	0.05	0.18
Supportive orientation	0.06	**0.20***	**0.26***	**0.39***	**0.27***	0.11
Rule orientation	0.02	0.05	0.08	-0.03	0.12	**0.22***
Meaningfulness of the job	-0.03	-0.05	-0.05	-0.02	-0.15	-0.01
Perceived responsibility	0.14	0.02	-0.08	-0.10	-0.23	-0.10
Conflict of function roles	-0.12	-0.02	0.00	-0.08	0.03	-0.05
Directive leadership	-0.06	-0.04	-0.04	0.02	0.07	-0.05
Entrusting leadership	0.02	0.04	0.04	0.03	-0.01	0.06
Variance remaining at level 2 (ICC)	5%	3%	4%	4%	7%	2%
ICC level 2 reduction compared to empty model	0%	2%	7%	5%	3%	4%
Intercept	3.73	2.30	1.22	2.64	2.51	1.88

Table 7. Six multilevel linear regression models. * p≤ .01 for level-I variables, p≤ .05 for level-II variables. Likelihood tests on model fit were all at a p ≤ 0.001 level. All coefficients are Beta-coefficients.

With respect to staff factors on level 2, two variables are found to be associated with perceptions of prison conditions. Obviously, the variables dealing with working with

inmates appear to be largely related. In housing units with staff who report a supportive orientation, inmates perceive various prison conditions more positively, such as *autonomy*, *monotony*, *program delivery*, and *contacts with the outside world*. This connection is in line with the expectation that a helpful staff facilitates desirable perceptions of prison conditions. Moreover, in housing units with a rule-oriented staff, inmates report that the rules and rights are clearer to them. There is also logic in this link: inmates acknowledge the rules when staff is clear about these rules. Evidently, both *helpful* and *rule enforcing* orientations of staff contribute to positive perceptions of prison conditions. The perception of safety is not connected to level 2 variables in the model. An explanation for that is the small amount of level 2 variance in the dependent variable (ICC is only 5%) and the fairly small standard deviation of the measurement scale (table 1, SD = 0,67). In comparison to the other measurement scales, the perceptions of safety are rather high ($\bar{x}$ = 3,45) which might indicate a leaning toward the unwillingness of inmates to show feelings of unsafety. Possibly, inmates do not want to show their feelings of safety to the public because it might make them fragile. An alternative explanation is that inmates feel relatively safe overall.

Scales concerning the topic of staff orientation show to be directly related to prison conditions. The other five staff scales (the human capital and leadership measures) are not significantly connected with inmates' perceptions of prison conditions. An explanation is that these topics may affect the well-being and correctional orientation of staff in a direct way, but are not directly related to perceptions of inmates.

In the empty models, there exists 5 to 11% of the variance in the dependent variables at the aggregate level (level 2). Except for the dependent measure of inmate safety, a considerable part of that variance at level 2 is explained by staff factors. In the measure of autonomy, 2% of the aggregate level variance is explained with the model, and 3% is still to be explained. Even seven out of eleven percent of the level 2 variance in the measure of monotony is explained. For program delivery, this is five out of nine percent. The model for contact with the outside world explained 3% of the 10% variance on the aggregate level. In conclusion, in the scale of rules and rights four out of six percent in total is explained on the second level.

With that, important information is generated on factors that are malleable when dealing with influencing perceived prison conditions of inmates. In the discussion we

raise some possibilities for factors which could be added to the models, and by which an even larger amount of level 2 variance might be explained.

5.6 Discussion

In this study, importation *and* deprivation factors are both used to come to grips with what determines the realization of some of the major goals of prisons (i.e. safety, humanity, and reintegration of inmates). It is assumed that, next to inmates' characteristics, environmental factors would affect these perceptions. Special attention is given to *staff* characteristics, orientation, and working circumstances that are believed to be important environmental factors. Importation factors as well as staff factors have demonstrated to be predictors of perceived prison conditions. The orientation of staff seems to be associated with an inmate's ability to satisfy his/her needs, such as realizing autonomy, having amenities, and engaging in activities. Therefore, the findings suggest that the importation and deprivation factors are complementary when perceived prison conditions are explained. Furthermore, the findings suggest that there are possibilities for a prison management to affect perceived prison conditions (and possibly inmate behavior resulting from those conditions), since staff factors are supposed to be malleable. Evidently, staff and management can help or hinder the satisfaction of the needs of inmates, such as the need for autonomy and activities. That is, these factors are malleable and contribute to the explanation of perceived prison conditions, next to less manageable factors (referred to in the importation theory), such as the age and ethnicity of inmates.

Relying on the analyses of coupled staff and inmate data, it is shown that staff's and inmates' perceptions of prison conditions are convergent and complementary to each other, which confirms the findings of prior research (Camp et al., 2002). Moreover, safety, human dignity, and efforts made on reintegration, as perceived by inmates, are connected to staff characteristics (i.e. staff orientation). Therefore, information stemming from surveys on inmates *and* staff has proven to be useful for a better understanding of the influences of prison conditions.

Furthermore, it is found that leadership and human capital measures are not (directly) related to the perceived prison conditions. However, in the present study it is

argued that the human capital factors may not influence inmate's perceptions directly, but can have repercussions on a staff member's attitude towards the correctional work. A case in point is the notion of Bottoms, who concluded in his synthesis on interpersonal prison violence and social order: The '[…] management of the prison can indeed indirectly affect prisoner-prisoner violence levels, there is a concomitant challenge to prison administrators to consider how they might best achieve reductions in prisoner-prisoner violence by thoughtful management changes' (Bottoms, 1999: p. 275). Apart from the specific case of prisoner-prisoner violence, the notion of Bottoms implicates that staff superiors may have an important *indirect* influence on prison conditions. For instance, superiors can incite staff to adopt a desirable orientation towards inmates. Offering education, courses and trainings are then common means. A pursuit of a balance in supportive and rule orientation could be considered since they both have positive connections with dimensions of perceived prison conditions. Furthermore, there are differences in perceived prison conditions stemming from maturity, prior incarcerations and ethnic background of inmates. For improving perceptions, a differentiated approach to these groups of inmates could be considered. Actually, more research is needed concerning how staff orientation towards inmates can be influenced.

5.7 Limitations of the Study

Some restrictions need to be addressed. First, the results are based on a cross-sectional data set, so we have no information about the causal pathways of our findings. Using a longitudinal design is difficult, because the inmate turnover in Dutch prisons is rather high[22]. Second, this study relies on self-report survey data (i.e. perceptions). Therefore, there is no certainty about the similarity between the *perceived* conditions and the *actual* prison conditions. However, prior research as well as the present chapter demonstrate that the perceptions of inmates vary consistently between prisons. Third, the fairly low response rates of the surveys are a weakness of the present study. Nonetheless, voluntary participation is guaranteed and the data sample represented the population. Fourth, the international comparability of the findings is somewhat limited since prison systems in other countries have a divergent classification of regimes.

22 In the Netherlands, the average time an inmate is incarcerated is about 3,5 months.

However, Dutch *remand prisons* may have considerable similarities with those of other European countries and American jails since they operate a limited program and house inmates who are waiting for trial. The Dutch *open regime* is somewhat comparable to low security facilities and the Dutch *prison regime* has substantial parallels with a medium security prison.

In the theoretical section, some factors suggested by human capital theory are assumed to be influential. In spite of those presumptions, these factors do not appear to be directly related to inmate perceptions. In future research, we believe that it will be necessary to make a clear distinction between human capital factors and prison staff orientation towards inmates. Hypothetically, we expect that perceptions such as the level of meaningfulness of the job, responsibility, and a conflict of function roles are determinants of staff orientation. Structural models can be considered to assess a possible connection with perceived prison conditions of inmates.

Further research should also focus on broadening the scope of data sources. The coupling of data is shown to be powerful when trying to explain complex phenomena. It also appears to be worthwhile to add a larger number of factors to the analysis of perceived prison conditions. For instance, inmates' neuropsychological functioning could be taken into account. In prior research, damaged cognitive functions accounted for the reaction of inmates to treatment in prison (Fishbein & Sheppard, 2006). Moreover, a shortcoming in the present study is the unavailability of behavioral measures. Researchers might make efforts to involve such measures for confirming or falsifying the presented results. Furthermore, environmental factors, like state employment and state crime rates, have proven to be influential in research on inner prison performance (Steiner, 2009). In addition to environment measures, future research may also include factors of aggregated inmate measures, such as the ethnical composition of the inmate population, the composition of staff, the collective criminal propensity (Camp et al., 2003), and the use of twin-bedded cells. Anyhow, there is still group variance to be explained in the measures of prison conditions presented in this study.

Chapter 6

Improving performance measurement: perspectives of stakeholders

In this chapter, a discussion of the research findings is provided using information obtained from eight interviews with relevant stakeholders active in the field of performance measurement in the Dutch prison system. By doing so, it is intended to answer the last sub-question of this book:

- According to relevant stakeholders, will the findings in this study contribute to accountability and performance improvement in the Dutch prison system?

Eight professionals were interviewed about their expectations that the research findings will contribute to accountability and performance improvement in the Dutch prison system. The interviews had no predetermined categories for answering the (open ended) questions[23] and the respondents were familiar with the methodological proposals formulated in this study. The conversations were held in the fall of 2013 in quiet meeting rooms (in a prison or at the headquarters of the national prison management) and lasted for one and a half to two hours. A voice recorder was used to record the exact narratives and all respondents agreed on the use of their narratives in this book. Two of the interviewees were government inspectors and six officials of the Dutch prison system (two prison directors, the former and the current head of the agency, the former head of information analysis and research, and the head of planning and control). In this way, the issues concerning performance measurement are viewed from different but relevant perspectives. The interest that they have in performance measurement, and the way they react to it (in terms of follow-up actions and measures[24]), varies between the respondents. We stress that the respondents are not a representative sample of all persons involved in doing or using performance measurement in the Dutch prison system. The aim of the interviews is to capture a variety of relevant views on the matter.

23 Questions that were posed to the interviewees included: 'What do you mean by performance measurements and what is the purpose of conducting these measurements?'; 'What are conditions, i.e. what are pros and cons, of performance measurement?'; 'Do you expect that the insights from the study contribute to the purposes of performance measurement?'; 'What are the impediments and resistances to the proper working of performance measurement after applying the proposed conditions?'

24 Some of them may use certain results of performance measurement for initiating extra inspections, others may pursue some organizations more closely, and others may develop plans to improve the performance result.

Paragraph 6.1 reports on the interviewees' narratives concerning their expectations of the effectiveness of an instrument for performance measurement using the conditions presented in this book. In paragraph 6. 2, the interview results are presented on future points of concern for performance measurement in the Dutch prison system.

6.1 On the expectations of the effectiveness of the instrument

The results of the interviews are organized in four blocks: 'Connection between mission and indicators'; 'Compare and contrast'; 'Multiple data collection methods'; and 'Detail of measurement'.

6.1.1 The connection between organizational mission and performance indicators

A major point of concern made by several interviewees is that the organizational mission and the current performance indicators used in the Dutch prison system are only partly related. As a former head of agency looks back to his years in charge:

> Actually, I never knew how the prisons factually were functioning. Every investigation and measurement highlighted different aspects, always fragmented. Every time I checked a performance result I thought: Is this a well-functioning facility? I missed the overall view to make a balanced assessment. I think it is possible to achieve such a view with much less indicators. At the time, two of my employees made about 38 indicators upon which I should steer; that is way too much. What is worse, this was developed without the field managers; how would we get any foundation for these indicators? It was everything but clear why and how the indicators should be measured; definitions were lacking. As a result, an instrument will never get to the substance of things. (Former head of agency)

The mismatch between the organizational mission and the indicators was one of the reasons why the Dutch Inspectorate of Security and Justice[25] made an exhaustive set of criteria for the assessment of the conditions in the prisons.

> Our framework for inspections is for many prison managers their 'implementation bible': the framework is clear; they know what they can expect. Such a comprehensive framework is lacking

[25] This inspection was then known as Inspectorate for the Implementation of Sanctions

> on the side of the national prison management to specify the desired performance. For example, the current planning and control cycle barely has qualitative elements. (Inspector 2)

The overview of tasks of Dutch prisons that is provided in this study (see Chapter 2) seems once again of importance. It is argued that such an overview helps to assess organizational performance more comprehensively. Furthermore, one respondent thinks that by improvements in performance measurement, the amount of audit and inspection activities can be reduced.

> The burden of inspection may decrease if the functioning of the performance system would improve. In any case, this system must cover all relevant areas of performance seen in light of the mission. In case an instrument for performance measurement functions adequately, we [the inspectorate and the national prison management] may work together and only inspect specific issues and do less integral inspections. An adequate performance system may help me to effectively introduce risk-based inspections. (Inspector 2)

Two conclusions can be drawn from these narratives. First, an assessment of the performance of Dutch prisons may be more comprehensive when the organizational mission and performance indicators coincide. Attention may be equally distributed to all elements of the organizational mission, which promotes a balanced performance measurement and management. Second, as performance will get measured more comprehensively, some respondents think that the implementation of prison sentences may be more strongly guided by the vision and performance indicators of the (national) prison management than by the criteria of the inspectorate. Maybe the inspection burden will decrease and the steering possibilities of the local prison managers will increase simultaneously.

6.1.2 Compare and contrast

Another topic that came up in the interviews is the background and application of performance measurement and performance comparisons in the Dutch prison system. According to several interviewees, these activities do not have their origin in goals of learning and improving.

> Benchmarks and rankings emerged from mistrust; the prison system was not centrally governed. At the time, the prison system grew in size and the amount of directors; the span of control became too large and all kind of lists and rankings were introduced to keep control [on the quality

> of implementation]. Recently, we try to rely more and more on trust between the prison manager and national headquarters, so we will need fewer indicators in future. (Prison director 2)

The mechanism of comparing and contrasting performance is recognized by the interviewees as a promising avenue to stimulate learning and improving.

> Lately, we gave a prison facility an award because of a certain best practice. This made a lasting impression in the prison, but outside the prison as well. Other prisons want to know what this best practice exactly comprised. (Inspector 1)

The mechanism of compare and contrast is potentially effective, but the comparisons should be fair and sound. That is, when measuring performance, we should account for 'given' circumstances that affect performance but cannot be influenced by the local prison managers. Rankings and benchmarks (i.e. activities that follow from performance measurement) are believed to be improved by the findings in this study.

> The new way [presented in the present study, TM] of measuring and visualizing performance creates much more reliability; there is no doubt about that. The way we used to compare prisons ran the risk of addressing the wrong prison director as the benchmark. Now there is an increased chance that we bring the right organizations together to learn from each other, because we account for their circumstances. (Former head of agency)

According to several interviewees, the proper working of the compare and contrast mechanism is not self-evident, and not (yet) part of everyday practice. However, they see promising signs, and the conditions presented in this study are thought to be supportive.

> Not all of the methodological proposals have been implemented, but I already notice differences in comparison with last year. Initially, rankings evoked resistance, but it has also led to alternative behavior and improved focus in management teams. [...] We should never directly judge the management on their performance scores. It all depends on the conditional and convenient circumstances, but the willingness to benchmark is present. Everybody wants to know where on the ranking list their organization is found. I believe that performance measurement is effective when our focus is on learning and improving and is not punitive. (Head of planning and control)

> Learning from each other seems to become more and more common. For example, in case of a calamity, an investigative committee of peer directors is mobilized afterwards. In the beginning, this led to resistance, but this has changed radically. (Inspector 1)

> Performance comparisons are an incentive for prison managements to improve performance. Solidarity seems to have returned, because local directors and national managers have more frequent dialogues, and there is less of a 'punitive performance culture.' (Inspector 2)

According to these narratives, the mechanism of comparing and contrasting performance seems to stimulate prisons managers to undertake action for improvement and cooperation between prisons, although this has not come about with ease. Notwithstanding the added value of the methods presented in this study, it is not enough, according to the interviewed professionals. They think that active involvement of several management levels, trust between those levels, and the absence of a 'punitive performance culture' are core ingredients for the proper working of performance measurement.

6.1.3 Multiple data collection methods

In several interviews, the issue came up how to measure performance comprehensively. There is no full agreement which data collection methods can be trusted, but the shared opinion is to thoroughly consider the meaning of every indicator.

> Some people label our periodical questionnaire among inmates as madness. But imagine this: if a specific activity is not appreciated by any inmate, one must be crazy to spend any more money on it. You need sources like that to make well-informed decisions. Indeed, some indicators do say nothing. The amount of complaints is pointed out to be a ridiculous indicator, because some individuals [inmates] make over a thousand complaints a year to undermine the daily management. Substantiated complaints may say more about prison performance, but I would always want to hear the story behind the numbers. (Former head of agency)

Several times the notion was brought up that one should not trust a single data collection method, because it may be biased and manipulated. Analyzing multiple data collection methods in conjunction would be better.

> Performance scores do not necessarily resemble the quality of the implementation of prison sentences. One can never entirely capture quality into a score. When you include all kinds of observations, like surveys, interviews, inspections and audits, you will get better measurement results, and differences between prisons can be uncovered. However, I have seen prisons with exemplary performance scores, while prison staff didn't communicate at all. (Inspector 2)

> Several sources are needed. You have to ask supplementary questions, taste the atmosphere, walk around. Every source gives some clue about performance, but one can only objectify the things when you visit the facility. (Inspector 1)

To sum up, every data collection method may give some information about aspects or dimensions of organizational performance. Several respondents argue that it must be thoroughly analyzed what the measurements exactly measure.

6.1.4 Detail of measurement

The interviewees do not fully agree on how detailed the performance information should be.

> We tend to capture the performance into absolute figures. Actually, my performance is not that I realize zero escapes. My results may be known from costs and efforts I needed to achieve zero escapes. The current performance indicators give limited qualitative information. If the prison climate is good, I know via staff and inmates whether or not there is tension in the prison, so I can take preventive measures. My performance may better be found in processes. (Prison manager 1)

It seems to be a complex decision as to what level of detail performance information should be presented. Most certainly, the interviewees agree that there should be a limited amount of indicators for reasons of interpretation. At the same time, careful and accurate monitoring is required.

> In the future, I intend to steer on some major principles. The prison directors are responsible for the local implementation. I don't need much detail afterwards from the control cycle because I am too late to steer on performance anyway. Instead of detailed performance measurement, I may have some of my people in each prison to monitor quality. They ensure that the Prison Act and other relevant laws and regulations are implemented accurately. Apart from that, it is the prison director's own business. I will govern every facility only on the same basic indicators; the methodological developments will help me to do that. However, we all remain a piece of the Ministry of Security and Justice, and should therefore always realize we are all representatives of the political leaders. They set the stage. But within those margins, I believe there is space enough to manage prisons in creative and fruitful ways. (Head of agency)

In any case, a balance between detail and user-friendly performance information seems the challenge for future improvements of performance measurement. Maybe composite indicators, as described in Chapter 4, will do justice to those concerns. A situation may be created in which the national prison management steers only on main principles (this promotes accountability), so that the 'let the manager manage' philosophy is adapted to

a greater extent (this promotes performance improvement).

6.1.5 Summing up

We learned, in general, that the interviewed stakeholders subscribe to the findings of the study. Applying the conditions articulated in this study, they expect that:

- performance will be measured more adequately and more comprehensively;
- a balanced focus on all elements of the organizational mission will stimulate performance, while the inspection burden for the prisons may decrease;
- performance comparisons will encourage managers to improve performance if there is managerial involvement and trust (including space for explanation of divergent performance scores and the absence of a punitive performance regime); and
- a balance between detail and user-friendly performance information will promote accountability and possibilities for performance improvement.

These narratives support our expectation that performance measurement – within the mentioned conditions – will lead to more accountability and performance improvement. In the next paragraph, we analyze more deeply some of the notions that came up in the interviews.

6.2 On future points of concern

After putting the conditions presented in this book into practice, it would be too optimistic to assume that performance measurement automatically leads to improvement. Therefore, supplementary questions were asked about future points of concern for effective performance measurement in the Dutch prison system. The answers of the interviewees fall into three categories of future points of concern: 'Effectuate a leaning culture'; 'The story behind the numbers'; and 'Character of the manager'.

6.2.1 Effectuate a learning culture

The interviewees stated that a learning culture is not easily created. Certain rules of conduct for managers may support such a culture. The interviewed inspectors are convinced that the findings in this study contribute to effective performance measurement; however, they think that more is needed to establish a learning culture.

> A measurement instrument using the proposed conditions provides in itself an improved insight into performance. But the next crucial thing is how the national prison management uses the instrument in its relation with the prison organizations. An active use will improve the performance, however, it is very counterproductive to only judge on performance scores; to achieve learning and improving, and to be accountable, a climate is needed in which a prison director can say: 'Hey, I need a little help with this.' You have to create a safe environment. (Inspector 1)

> You'd better not initially judge on performance information; there must be a learning curve. If this process is not set in motion, the national prison management can of course intervene. The instrument can be used to address people, but they must get the chance to take action. When you don't see improvements over the course of time, then you may have a conversation [between national prison management and the local prison manager]. One should have open discussions because a patronizing attitude will be demotivating. (Inspector 2)

The interviewees state that managers should actively use results of performance measurements. In case of negative results, for example, a manager develops a plan to improve performance. At the same time, a safe managerial environment may promote learning activities and cooperation between the organizations. When this still does not lead to actions and performance improvement, according to the interviewed stakeholders, national prison management may steer and intervene. However, as was stated in the first chapter, the national prison management has no absolute power over all organizations for which they are responsible. Since (semi-)public organizations, (including prisons) are placed at arm's length of central government they may implement their tasks in their own way.

> How to create a climate of learning? That is difficult, especially because some of my facilities are privatized. Competition may stimulate performance, but it may also create a situation where the organizations do not want to share their experiences and best practices. When I would have one or two representatives in every facility, they may steer towards learning. And the inspectorate can play an important role as well. For example, they may discuss their findings with all layers in the organization. (Head of agency)

This may ensure a learning culture and the active use of the results of performance measurement. However, it cannot be predicted how managers will react to the (permanent) presence of assistants of the national prison management and the active interference of inspections.

As mentioned in the interviews, another factor that may contribute to a culture of learning is 'calmness' within the organization. In other words, a culture of learning will benefit from circumstances that are not too tempestuous.

> Leadership is crucial. When it comes to quality of the implementation, the planning and control instrument is not activating in itself. For quality improvement, leadership and governance is needed. It lacks decisiveness in several managerial layers, and the prison system faces too much policy changes. In crucial occasions, managers are replaced too quickly, so their measures cannot 'work out'. (Inspector 2)

6.2.2 The story behind the numbers

Future users of performance measurement should expound on the context of the measured results because not everything can get captured with measurements. An example of that statement is given by a prison manager regarding the negative effects of utilizing the instrument of performance measurement, and the resulting comparative graphs not being presented in a nuanced way.

> Images of an organization are hard to change, and performance graphs require explanation. Definitions need to be clear. Goals should be achievable and, as a consequence, cannot be the same for all organizations. Nothing is more demotivating than goals that cannot be achieved. Prisons differ with respect to their circumstances, and may therefore have various targets. Some prisons cannot realize a balanced budget because they have older personnel, and that is something you cannot change over one year or even over ten years. (Prison manager 1)

This narrative gives an indication why performance measurement should go along with managerial explanation, such as in texts incorporated in an annual report or in face-to-face conversations in which managers provide accounts to higher level managers. Another prison manager also warns for too easy conclusions derived from performance measurements.

> We used to have quick opinions about good or bad. That didn't work; we need to discuss what it is that produces bad performance scores. We need a culture that those things are debatable. (Prison manager 2)

An example is provided by a former head of agency who wanted to ensure he was 'in control' of the implementation in all prisons, and who had to rely on output measurements and conversations he had with the local prison managers.

> Transparency can force organizations to decrease performance differences. But at the national level of a large agency like the prison system, there needs to be trust. A good prison manager gives me information to steer. A relationship based on mutual trust is essential, but in practice it is exceptional. A prime example is a prison manager that once said to me that there hadn't been any escapes over the last fifteen years from his prison. Six months later an inmate escaped from that very prison, and my audit team discovered that the prison was as leaky as a sieve. The performance indicator about escapes therefore only gives limited information about the security situation of a prison. You don't know how well the work processes are arranged to prevent escapes. A fixation on output creates a situation in which you are too late to change course. 'The process' is often forgotten and I think that in that process there is a world to win. (Former head of agency)

6.2.3 The character of the manager

The role of the local prison manager is often expressed in the interviews. The interviewed prison managers are asked to reflect on the reasons why colleagues are (not) open to learn from each other in order to improve their performance.

> The willingness to learn from each other is weak among Dutch prison managers. That is partly caused by egoism, and partly caused by the fact that the managerial environment is not safe for them. Yet, the sticking point actually lies in the personality of the prison managers. People that are open to learning and improving obviously do not easily rise to management positions. (Prison manager 1)

It must be noted that other interviewees argued that the willingness to learn has evolved in recent years.

> In the old days, managers were mainly concerned with 'their product' and worked purely for the inmates. This involvement [with inmates] is here to stay, but financial and economic management has been strengthened, as well as the willingness to learn from each other. Good management translates to a stable prison and a good prison climate for inmates. […] The new type of manager is more approachable and wants to give account of his or her performance. That is the development that is going on in the Dutch prison system. Previously, managers shielded their

> performance so that headquarters couldn't see their real performance. Happily, those days are gone; although visions of prison managers [on the implementation of prison sentences] still differ. Some managers see the norms of the Dutch Prison Act as the minimum norms; others interpret them as the maximum. This difference in interpretation produces differences in implementation. (Prison manager 2)

Prison managers may be open to be addressed on their performance (e.g. by national prison management or the media), but the discussion will hold back when there is no agreement about which tasks they are expected to perform.

6.2.4 Summing up

In this paragraph we used the narratives of eight respondents that are involved in the measurement of performance in the Dutch prison system. Although these respondents are no representative sample of all stakeholders in the Dutch prison system, the interviews show a variety of views on what is needed for performance measurement in the (near) future to contribute to accountability and performance improvement.

Firstly, the interviewees think that a learning culture is not supported by a punitive performance regime. In contrast, they state that performance differences should be made debatable, and there must be tolerance for diverging performance scores. This tolerance has - of course - its limits; when improvement does not take place in the longer run, some respondents argue, national prison management may intervene. Secondly, interviewees stated that performance graphs and ranking lists may create an image of an organization that is difficult to change; performance scores should therefore be accompanied with clarifications of the management. Therefore, the interviewees believe that the national prison management and local managers should cooperate in a spirit of mutual understanding. Evidently, not everything can be expressed with (numerical) performance measurements. Thirdly, one respondent stated that some prison managers shielded or hided their (poor) performance. However, another respondent reported that this is changing in the Dutch prison system in recent years. In any case, transparency is not self-evident and several interviewees stated that there must be faith that a final judgment about performance scores will not be rushed.

Chapter 7

General conclusions

In the first paragraph of this final chapter, we recall the aims of the study. Next, we formulate answers to the research questions. Finally, the limitations and possibilities for broader application of the research findings are discussed.

7.1 Aims of the book

This study started with the notion that (semi-)public organizations in Western countries have been placed at arm's length of central government during the last three decades. While allowing these organizations more space for implementing tasks in their own way, central governments and their ministers have stayed responsible for the organizational achievements, and therefore want to monitor their performance. One of the key instruments that are used in this respect is *performance measurement*. Next to accountability reasons, performance measurement is used for the improvement of the performance of (semi-)public sector organizations. By comparing and contrasting performance scores, organizations are urged to learn from each other and improve their performance.

Besides the well-known advantageous effects, more and more evidence shows that performance measurement also elicits negative side effects. An example of such an effect is that organizations hide their true performance by 'gaming' the numbers in order to 'look good' to central government. Other examples are tunnel vision and myopia, where organizations only focus on what is measured, and prefer to give attention to short-term goals. In this study, we focus on methodological conditions for performance measurement to prevent negative side effects and promote performance improvement in the context of the Dutch prison system.

In Chapter 1, several methodological notions are described conditional for the proper working of performance measurement. The first condition is to have a clear picture of the goals and tasks that are applicable to the sector upon which the performance measurement is focused. This is needed to measure performance in a valid way. The next condition concerns the reliability and validity of data collection methods. To rely on one data source may be risky, because measurement biases can lead to a distorted view. The use of multiple data collection methods may reduce these risks and lead to a more

comprehensive view of performance as well. This study therefore seeks for an explicit measurement strategy making use of a combination of measurement methods.

A further condition is that circumstantial factors a local prison management *cannot* influence are taken into account when measuring performance. Since these circumstances (1) are 'given'; (2) may affect performance scores; and (3) can be quite different for different organizations confronted with performance measurement and comparisons, methodological and statistical procedures can facilitate an apples-to-apples comparison. A final condition of performance measurement discussed in this study concerns the probability that a local prison management *can* influence a performance score. Managerial actions are ideally based on scientific evidence about how organizational performance can be influenced. This type of knowledge can promote the 'learn and improve function' of performance instruments.

The research questions of this book are derived from the above mentioned conditions for performance measurement in the specific situation of the Dutch prison system.

7.2 Summary of research findings

This paragraph summarizes the research findings presented in this book.

7.2.1 A matter of balance: about the abstract goals of imprisonment and task ambiguity in Dutch prisons (Chapter 2)

Chapter 2 describes the tasks of the Dutch prison system derived from the key goals of imprisonment. The research question is the following:

- Following the goals of imprisonment, what are the tasks of Dutch prisons?

The key goals of imprisonment are largely clear, since most prisons should provide safe, humane, and rehabilitative conditions. But if we look more closely, complex ambiguities seem to be at work. In this chapter, the specific situation of the Dutch prison system is studied on how implementation differences can arise within a seemingly detailed regulatory framework. It is concluded that on certain parts of the organizational

mission, a clear description of the pursued performance is available, mostly linked to the Dutch Prison Act and related regulations. Examples are the standardization of airing (at least one hour per day) and visiting hours (at least one hour per week). On other parts, laws and regulations give less guidance.

Two core penological principles of imprisonment in the Netherlands – namely, the principles of minimal restrictions and rehabilitation – can be interpreted and applied in quite different ways. Regarding minimal restrictions, the law stipulates that the restrictions imposed on inmates should be reduced to a minimum, as long as the security in the prison and the goals of imprisonment are not called into question. Subsequently, 'the goals of imprisonment' are not specified in the law, and the relationship between 'the security of the prison' and currently imposed restrictions is anything but self-evident. We conclude that it is not always precisely clear for what tasks Dutch prisons are deemed responsible and performance may therefore not be measured comprehensively.

7.2.2 A method to deal with dissimilar circumstances of public organizations in performance comparisons: evidence from Dutch prisons (Chapter 3)

Performance measurements may give a view of differences in goal accomplishment between organizations and thus basically allow for comparing organizations. However, organizations may operate in quite different 'given' circumstances which may affect their performance. Apples-to-apples comparisons are therefore not self-evident, which leads to the following sub-question, in which we –again – concentrate on the Dutch prison system:

- Do Dutch prisons have equal contextual circumstances, and if they do not, how can these differences be dealt with when making performance comparisons?

When performance scores are not fully related to the efforts of an organizational management, it is argued to make statistical adjustments for factors that the management cannot control. If we would omit such notions, there is every chance that the wrong benchmark is indicated, thereby leading to organizations exchanging practices that may not contribute to performance improvement. It is known from

several fields (e.g. hospitals, prisons, schools) that adjusting performance scores changes the rankings of the organizations. To date, the added value of performance score adjustment has been shown in different fields, however, their use is anything but common practice. It is therefore argued in this study that a systematic step-by-step plan for performance comparisons is needed, which must include such adjustment techniques.

First, the reliability, validity, and variability of performance measurements must be ensured. Second, the plan prescribes to examine factors that influence performance measurement. Three types of factors are distinguished: non-discretionary factors (factors that are beyond the influence of organizational management); valid variance (factors that are within the control of management); and random variance (the disturbance term of a stochastic model). To assess which factors are expected to be related to the performance measurement, as well as those that are outside the sphere of managerial influence, we used an expert meeting. Third, we proposed that the last mentioned factors are taken into account by including them in a regression model that predicts the expected score for every organization within the comparison. Thus the difference between the predicted and the realized score can be expressed into a so-called empirical Bayes residual, which is interpreted as the part of performance that can be influenced by the management. The step-by-step plan finally prescribes to rank the organizations on their residuals. When these steps are followed, a contribution is made to 'level the playing field' in favor of sound and fair performance comparisons.

7.2.3 Measuring performance in the public sector: Towards a measurement strategy for composite indicators (Chapter 4)

Performance of (semi-)public sector organizations often concerns a complex constellation of phenomena that can be perceived quite differently by various stakeholders. As a consequence, (semi-)public sector performance is not always easily measured.

- Which measurement strategy may comprehensively assess organizational performance, and give an account of limitations of various data collection methods?

It is argued that the multifaceted nature of (semi-)public sector performance induces the application of multiple data collection methods. The use of multiple measurements will not prevent us from making errors in measurement, but it is likely that the weaknesses of certain methods may be counterbalanced by the strengths of others. In order to reduce complexity and to provide user-friendly performance information, the different measurements may be taken together in *composite scores*.

In this book, critical realism is adopted as an appropriate epistemological background for the assessment of (semi-)public sector performance. The approach simultaneously takes the (performance) phenomena, the underlying mechanisms, and their contextual factors into account. In case several measurements derived from different data collection methods concerning the same performance theme (hereafter called 'dimension') converge, we may conclude that a certain part of organizational performance is adequately assessed. Furthermore, the measurement strategy also offers guidance when measurements in the theorized dimension show divergence. In case reliability and validity are established in two measurements that diverge, one measurement may have a better fit in an adjacent performance dimension. To establish a comprehensive measurement of a performance dimension, we are looking for convergence that is indicated by statistically significant and relevant correlates with the expected sign (positive or negative). However, we do not want to have too strong correlates, as highly correlated measurements in a performance construct run the risk that they measure the same performance aspect. Also, high correlates may lead to an overemphasis of a particular aspect when we make up a composite score. Therefore, we may comprehensively assess performance, when the theorized factors in our performance construct – including outcomes, mechanisms, and context – are measured with *multiple data collection methods and* are *significantly but moderately related* with the *expected sign*.

To empirically test whether this measurement strategy gives guidance to the performance assessment of (semi-)public organizations, the strategy is applied to the case of *staff safety in the Dutch prison system*. The empirical findings support the theorized dimension of prison staff safety. However, some measurements concerning

the context we theorized to be part of the performance dimension, are shown to be unrelated to the outcome measurement. These findings gave rise to a reconsideration of the validity of the theorized dimension in future research. Nevertheless, we found that the building blocks of the comprehensive measurement of staff safety can be derived from different data collection methods. This supports the proposed measurement strategy, and it is therefore argued that the strategy is a promising avenue for the development of comprehensive assessment of multifaceted performance in the broader field of (semi-)public services.

7.2.4 The influence of prison staff on inmate conditions: A multilevel approach to staff and inmate surveys (Chapter 5)

In this chapter, it is analyzed whether measurements of performance are related to factors that can be influenced by the management of a (semi-)public sector organization. This knowledge can be of importance when organizations pursue improvements to their performance. Here we focus on some major goals of the Dutch prison system, which mainly concern the prison conditions for inmates. The conditions as perceived by inmates are adopted as measurements of prison performance.

- Do performance measurements relate to factors that can be influenced by a local prison management?

A hypothetical model was developed based on two theories about the underlying influences on perceptions of prison conditions. Import theory seeks for causes of inmate perception and behavior in the history and characteristics of the inmates themselves. Factors like the inmates' experiences in their youth, prior offences, and personality (disorders) are thought to be of influence on inmates' perceptions and behavior. The second theory used, the deprivation theory, explains perceptions and behavior of inmates in relation to their dissatisfaction; that is, they are deprived of liberty, security, autonomy, amenities, and so on. These 'pains of imprisonment' can be mitigated by certain circumstances that can be created in a prison context; for example, positive contacts with staff and the application of the house rules.

Using the results of simultaneously held surveys among staff and inmates, it is found that in housing units where the orientation of staff towards inmates is relatively supportive, inmates perceive their conditions as more positive. Furthermore, inmates report that they find the rules and rights of the prison relatively clear when staff are oriented on enforcing house rules and emphasizing the structure of the daily program of inmates. These factors – that may be influenced by local prison management – are shown to be relevant, even when we control for inmate characteristics (such as age and ethnicity). The findings suggest that the importation and deprivation factors are complementary when perceived prison conditions are explained. Therefore, a local prison management has opportunities to influence performance, at least in cases where performance is affected by perceived prison conditions.

7.2.5 Expectations for implementation of the conditions in the Dutch prison system (chapter 6)

In the interviews, eight stakeholders of performance measurement in the Dutch prison system subscribed to the findings of the study, and think that the methods – if well applied – contribute to accurate performance measurement.

The interviewed stakeholders also expressed that the conditions described in this book may not be enough to guarantee an instrument for performance measurement that contributes to accountability and performance improvement. The interviewees expect, for example, that the managerial climate and performance regime should not be punitive, and that there should be tolerance for performance differences between organizations.

7.2.6 Answering the central research question

The central research question reads as follows:

- How may conditions for accurate measurement of performance be applied to the Dutch prison system and may performance measurement within those conditions lead to an increased likelihood of accountability and performance improvement in Dutch prisons?

The sub-questions guided the search for conditions of accurate performance measurement in the Dutch prison system. The findings resulting from the sub-questions are used for answering the main research question.

This study provided an overview of the tasks that Dutch prisons have to perform. At many points laws and regulations clarify the tasks for which prisons are deemed responsible, although there is a lack of clarity at some points. Once a clear overview of the tasks can be provided, performance can get measured comprehensively. Performance indicators may then give a balanced view of the efforts made to achieve the organizational mission. When a set of performance indicators reflects (all elements of) the mission, the chance of perverse effects (like tunnel vision) may be smaller. That is important because perverse effects may interfere with accountability and performance improvement. To date, this condition of performance measurement is partly at hand in the Dutch prison system.

Furthermore, a step-by-step method is provided to compare performance of organizations while accounting for the diverging 'given' circumstances. The exercise of 'leveling the playing field' has proved to be meaningful in the Dutch prison system. Performance comparisons therefore may lead to the matching of the right benchmark partners (superior performers can 'teach' and inferior performers can learn) to exchange experiences in order to improve performance.

Next, accurate performance measurement ensures the comprehensive assessment of the sometimes complex and multifaceted goals that (semi-)public services strive for. In this study, we get to grips with the condition of having a measurement strategy: we assessed performance by including outcomes, mechanisms, and context measurements derived from several data collection methods. Together these measurements may comprehensively measure a specific (and multifaceted) task of Dutch prisons. After testing these considerations on empirical data concerning prison staff safety, it is concluded that performance phenomena should be modeled in a network and incrementally adjusted to reach a model in which the measurements are significantly but moderately related with the expected sign. In this way we may ensure that the measured elements are indeed related but also measure different parts of the same performance phenomena (so no element is overemphasized).

Furthermore, the current study found evidence that performance measurements are related to factors a local prison management can influence. If that would not be the case, performance measurement may have a perverse impact on performance because managers cannot improve their performance scores and get demotivated.

Overall, we found possibilities and developed methodological solutions to put the conditions of accurate performance measurement into practice in the Dutch prison system. Furthermore, eight respondents expressed in interviews that they expect the conditions described in this book to contribute to accountability and performance improvement. Simultaneously, these interviewees expect that such conditions may help prevent perverse effects of performance measurement as well. However, their assessment is that there should be 'managerial conditions' as well. For example, the national prison management and the local prison managers should uphold a non-punitive performance regimen and a managerial culture in which managers feel safe to ask for help to improve their performance.

7.3 Limitations and broader application of the research findings

Specific limitations of the research findings are described in the subsequent chapters. Therefore, we restrict ourselves to the general limitations of the book.

A limitation of this book is that we used data of the fiscal years 2006–2007. More recent data, of a similar breadth, depth and width, were not available (when we started the study). However, we argue that the data are not outdated for the purpose of this book, since all of the used databases are still operational to date. Moreover, the focus is on methodological development and not on performance judgments as such.

With regard to the narratives presented in chapter 6, some limitations should also be mentioned. We argue that the evidence that these narratives generates is modest because the respondents are no representative sample and only eight interviews were conducted. The conversations only give an idea of the thoughts and views of some relevant professionals involved in the performance measurement of the Dutch prison system.

A further limitation of the study is the use of theory. We benefitted from existing theoretical insights (e.g. derived from penology, philosophy, public administration,

sociology, and statistics) and applied these when we took up the methodological challenges. However, since the data-collection (already existing data from registers, surveys and audits) was not based on the theories used, the testing of theories was limited to the fifth chapter on importation and deprivation theory.

Next to these general limitations, we also point out opportunities for broader application of the research findings.

This study was limited to the Dutch prison system – a public service that certainly has its own signature, organizational structure, and culture. However, the agency is also part of a larger family of *total institutions* that also strive for realizing complex goals, and which may therefore benefit from the conditions of effective performance measurement presented here. These organizations may therefore invest in thorough analysis, for example, to bridge the gap between their mission and performance measurements, to detect valid and reliable data collection methods, and to build a comprehensive measurement strategy in order to assess (well-defined) performance dimensions. In several fields of (semi-)public service – be it in the Netherlands and beyond – many organizations already meet *some* of those conditions. However if these conditions are applied simultaneously, effective performance measurement may be supported more strongly.

Although the application of the conditions distinguished in this book is promising, (semi-)public services should be conscious of what is needed to make performance measurement a successful activity. Evidently, measuring performance is a labor-intensive activity if we want to ensure that it contributes to improving – and giving account of – performance. The performance of (semi-) public services is not easily measured and visualized for the purpose of interpretation. Among the unavoidable pitfalls of performance measurement is the problem that every (set of) measurement(s) is – in the end – only an approximation of (semi-)public sector performance. Certain measurements may be substantially biased, or may reflect parts of a different performance dimension than was intended to assess.

Other points of attention include the relation between costs and benefits of performance measurement, validity and reliability of measurements, the availability of performance data, proper use by managers, and straightforward interpretation. In other organizations, there may not be a measurement tradition similar to the Dutch prison system. To start from the bottom up, developing and implementing an instrument for performance measurement may take considerable efforts. The methodological guidance provided in this study can be of assistance.

References

A

Arnold, H., Liebling, A., & Tait, S. (2008). Prison officers and prison culture. In Y. Jewkins (Ed.), *Handbook on prisons* (pp. 471-495). Uffculme: Willan Publishing.

Aronson, E. (2001). Integrating leadership styles and ethical perspectives. *Canadian Journal of Administrative Sciences*, *18*(4), 244-256.

Astbury, B., & Leeuw, F.L. (2010). Unpacking black boxes: mechanisms and theory building in evaluation. *American Journal of Evaluation*, *31*(3), 363-381.

B

Bales, W.D., Bedard, L.E., Quinn, S.T., Ensley, D.T., & Holley, G.P. (2005). Recidivism of public and private state prison inmates in Florida. *Criminology and Public Policy, 4*(10), 101-127.

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.

Bass, B.M., & Steidlmeier, P. (1999). Ethics, character, and authentic transformational leadership behavior. *The Leadership Quarterly, 10*, 181-117.

Bechara, J., & Van de Ven, A. H. (2011). Triangulating philosophies of science to understand complex organizational and managerial problems. *Research in the Sociology of Organizations*, *32*, 343-364.

Behn, R.D. (2003). Why measure performance? Different purposes require different measures. *Public administration review*, *63*(5), 586-606.

Bentham, J. (1995) Panopticon. In M. Bozovic (Ed.), *The Panopticon Writings* (pp. 29-95). London: Verso.

Berman, B.J.C., Brenman, J., & Vasquez, V. (2010). Using trained observer ratings. *Handbook of Practical Program Evaluation*, 298-320.

Bevan, G. (2006). Setting targets for health care performance: lessons from a case study of the English NHS. *National Institute Economic Review*, *197*(1), 67-79.

Bhaskar, R. (1978). *A realist philosophy of science*. Harvester Wheatsheaf: Hemel Hempstead.
Bhaskar, R. (2008). *A realist theory of science*. Taylor & Francis US.
Biessen, P.G.A. (1992). *Oog voor de menselijke factor; achtergrond, constructie en validering van de basisvragenlijst Amsterdam*. Lisse: Swets & Zeitlinger.
Biessen, P.G.A., & De Gilder, D. (1993). *BASAM: Basisvragenlijst Amsterdam: Handleiding.* Lisse: Swets & Zeitlinger.
Bilby, C. (2008). Does it really matter what offenders think? The importance of uncovering offenders' experiences in prison and on probation. *Prison Service Journal, 177*, 38-42.
Birgden, A. (2004). Therapeutic jurisprudence and responsivity: Finding the will and the way in offender rehabilitation. *Psychology, Crime &Law*, *10*(3), 283-295.
Blad, J. (2003). *Elementen van een herstelgericht detentieregime*. Rotterdam: Erasmus Universiteit.
Bloom, N., Propper, C., Seiler, S., & Van Reenen, J. (2010). *The impact of competition on management quality: evidence from public hospitals.* (No. w16032). National Bureau of Economic Research.
Bogaerts, S. & Den Hartogh, V. (2008). *Onderlinge agressie en geweld van personeelsleden in een penitentiaire inrichting*. Den Haag: WODC.
Boin, A. (2001). *Crafting public institutions: Leadership in two prison systems*. LynneRienner Publishers.
Boone, M.M. (2000). *Recht voor commuun gestraften*. Deventer/Gouda: Quint.
Boone, M.M. (2007). Selective rehabilitation. In M.M. Boone & M. Moerings (Eds.), *Dutch Prisons* (pp. 231-249). Den Haag: Boom JuridischeUitgevers.
Bottoms, A.E. (1999). Interpersonal violence and social order in prisons. *Crime & Justice, 26*, 205-281.
Bryman, A. (2012). *Social research methods*. Oxford university press.
Burns, J.M. (1978). *Leadership*. New York: Harper & Row.

C

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.

Cameron, W.B. (1963). *Informal sociology: A casual introduction to sociological thinking*. Vol. 21. New York: Random House.

Camp, R.C. (1989). *Benchmarking: The Search for Industry best practices that lead to superior performance.* ASQ Quality Press: Milwaukee.

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.

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.

Camp, S.D., Gaes, G.G., Langan, N.P., & Saylor, W.G. (2003). The influence of prisons on inmate misconduct: a multilevel investigation. *Justice Quarterly*, *20*(3), 501-533.

Camp, S.D. & Gaes, G.G. (2005). Criminogenic effects of the prison environment on inmate behaviour: some experimental evidence. *Crime & Delinquency, 51*(3): 425-442.

Campbell, D.T. (1979). Assessing the impact of planned social change. *Evaluation and program planning*, *2*(1), 67-90.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81-105.

Cao, L., Zhao, J. & Van Dine, S. (1997). Prison disciplinary tickets: a test of the deprivation and importation models. *Journal of Criminal Justice, 25*, 103-13.

Carnap, R. (1956). The methodological character of theoretical concepts. *Minnesota studies in the philosophy of science*, *1*, 38-76.

Charnes, A., Cooper, W.W. & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*, 429-444.

Cheeseman, K.E., Mullings, J.L., & Marquart, J.W. (2001). Inmate perceptions of security staff across various custody levels. *Corrections Management Quarterly*, *5*(2), 41-48.

Clemmer, D. (1940). *The prison community*. New York: Holt, Rinehart & Winston.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciencies*. Routledge.

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Council of Europe (2006). *European Prison Rules*. Rec(2006)2.

Craig, S.C. (2004). Rehabilitation versus control: an organizational theory of prison management. *The Prison Journal, 84*(4), 92S-114S.

Cronbach, L.J. (1947). Test reliability: Its meaning and determination. *Psychometrika*, *12*, 1-16.
Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281-302.

D

Daggett, D.M. & Camp, S.D. (2009). Do official misconduct data tell the same story as the individuals who live in prison? *Criminal Justice Review*, *35*, 200-219.
De Keijser, J.W. (2004). Doelen van straf. Morele theorieën als grondslag voor een legitieme strafrechtspleging. In B. Van Stokkom (Ed.), *Straf en herstel, ethische reflecties van sanctiedoeleinden* (pp. 43-65). Den Haag: Boom Juridisch Uitgevers.
De Leeuw, E.D. (2005). To mix or not to mix data collection modes in surveys. *Journal of official statistics*, *21*(5), 233-255.
De Wolf, I.F. & Janssens, F.J. (2007). Effects and side effects of inspections and accountability in education: an overview of empirical studies. *Oxford Review of Education*, *33*(3), 379-396.
DiIulio, J.J. (1987). *Governing prisons; a comparative study of correctional management.* New York: Free Press.
Dillman, D.A., Smyth, J.D. & Christian, L.M. (2009). *Internet, mail and mixed-mode surveys*. Hoboken, NJ: Wiley.
Dillman, D.A., & Tarnai, J. (1988). Administrative issues in mixed mode surveys. *Telephone survey methodology*, 509-528.
DJI [Dutch Correctional Agency] (2004). *Gedetineerd in Nederland 2004*. [Being remanded in the Netherlands 2004].
DJI [Dutch Correctional Agency](2009) *Strategisch kader Modernisering Gevangeniswezen*. Den Haag: Programma MGW: The Hague.
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.

E

Edwards, J. R. (2003). Construct validation in organizational behavior research. *Organizational behavior: The state of the science*, 327-371.

Eisenkopf, G. (2009). Negative weights for performance measures. *International Public Management Journal*, *12*(3), 332-344.

Evans, J. R. (2004). An exploratory study of performance measurement systems and relationships with performance results. *Journal of Operations Management*, *22*(3), 219-232.

F

Farkas, M.A. (1999). Correctional officer attitudes toward inmates and working with inmates in a "get tough" era. *Journal of Criminal Justice, 27*(6), 495-506.

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Fishbein, D., & Sheppard, M. (2006). *Assessing the Role of Neuropsychological Functioning in Inmates' Treatment Response*. Report submitted to the U.S. Department of Justice.

Franke, H. (1995). *The emancipation of prisoners. A socio-historical analysis of the Dutch prison experience*. Edinburgh: Edinburgh University Press.

Fried, H.O., Lovell, C.A.K., Schmidt, S.S. & Yaisawarng, S. (2002). Accounting for environmental effects and statistical noise in Data Envelopment Analysis. *Journal of Productivity Analysis, 17*, 157-174.

Fryer, K., Antony, J., & Ogden, S. (2009). Performance management in the public sector. *International Journal of Public Sector Management*, *22*(6), 478-498.

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.

G

Gaes, G.G., Camp, S.D., Nelson, J.B., & Saylor, W.G. (2004). *Measuring prison performance, government privatization & accountability*. Walnut Creek, California: Altamira Press.

Gaes, G.G. & Camp, S.D. (2009). Unintended consequences: Experimental evidence for the criminogenic effect of prison security level placement on post-release recidivism. *Journal of Experimental Criminology, 5*, 139-162.

Garland, D. (1990). *Punishment and modern society*: *a study in social theory*. Chicago: University of Chicago Press.
Garland, D. (2001). *The Culture of Control: Crime and Social Order in Contemporary Society*. Chicago: University of Chicago Press.
Gendreau, P., Goggin, C.E., & Law, M.A. (1997). Predicting prison misconducts. *Criminal Justice and Behavior*, *24*(4), 414-431.
Gianakis, G.A. (2002). The promise of public sector performance measurement: anodyne or placebo? *Public Administration Quarterly*, *26*(1/2), 35-64.
Goffman, E. (1961). *Asylums: Essays on the social situation of mental patients and other inmates*. Chicago: Aldine.
Goldstein, H., & Spiegelhalter, D.J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*(3), 385-443.
Grapendaal, M. (1990). The inmate subculture in Dutch prisons. *British Journal of Criminology, 20*(3), 341-357.

H

Hagedoorn, J., & Cloodt, M. (2003). Measuring innovative performance: is there an advantage in using multiple indicators?. *Research policy*, *32*(8), 1365-1379.
Halachmi, A. (2005). Performance measurement is only one way of managing performance. *International Journal of Productivity and Performance Management*, *54*(7), 502-516.
Hamann, P. M., Schiemann, F., Bellora, L., & Guenther, T. W. (2013). Exploring the dimensions of organizational performance a construct validity study. *Organizational Research Methods*, *16*(1), 67-87.
Harer, M., & Steffensmeier, D. (1996). Race and prison violence. *Criminology*, *34*, 323-355.
Hatry, H.P. (2010). Using agency records. In J.S. Wholey, H.P. Hatry & K.E. Newcomer (Eds.), *Handbook of practical program evaluation*(pp. 243-261). New York: Jossey-Bass/Wiley.
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.

Hemmens, C., & Stohr, M.K. (2001). Correctional staff attitudes regarding the use of force in corrections. *Corrections Management Quarterly*, *5*(2), 27-40.
Hobbs, G.S., & Dear, G.E. (2000). Prisoners' perceptions of prison officers as sources of support. *Journal of Offender Rehabilitation*, *31*(1/2), 127-142.
Hochstedler, A. & DeLisi, M. (2005). Importation, deprivation, and varieties of serving time: An integrated-lifestyle-exposure model of prison offending. *Journal of Criminal Justice, 33*, 257-266.
Hoekendijk, G.P. & Kommer, M.M. (2011). Strafdoelen en tenuitvoerlegging: perspectief op een nieuwe verenigingstheorie?. *Sancties,* 4: 212-222.
Hood, C. (1995). Contemporary public management: a new global paradigm?. *Public policy and administration*, *10*(2), 104-117.
Hood, C. (1991). A public management for all seasons?. *Public administration*, *69*(1), 3-19.
Howell, J. M., & Higgins, C. A. (1990). Champions of technological innovation. *Administrative science quarterly*, *35*(2), 317-341.
Hox, J.J. (2010). *Multilevel Analysis: Techniques and applications* (2nd ed.). New York: Routledge.
Huey Dye, M. (2010). Deprivation, importation, and prison suicide: Combined effects of institutional conditions and inmate composition. *Journal of Criminal Justice, 38*, 796-806.

I

Inspectorate for the implementation of sactions (2009). *Inspectiejaarbericht 2008*. Den Haag: Ministerie van Justitie.
Irwin, J., & Cressey, D. (1962). Thieves, convicts, and the inmate culture. *Social Problems, 10*, 142-155.

J

Johnson, R.B., & Gray, R. (2010). A history of philosophical and theoretical issues for mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research* (pp. 69-94). Thousand Oaks: Sage.
Jonkers, W.H.A. (1975). De strafrechtelijke straf: inhoud, grondslag, doeleinden. In *Het Penitentiair recht* (pp. 1-22), Losbladig, p. III.

K

Kelly, J. M., & Swindell, D. (2002). A Multiple-Indicator Approach to Municipal Service Evaluation: Correlating Performance Measurement and Citizen Satisfaction across Jurisdictions. *Public Administration Review*, *62*(5), 610-621.

Kelk, C. (2010). *Materieel Strafrecht.* Deventer: Kluwer.

Kempe, G.Th. (1973). Franz von Liszt en de criminologie. In *Mededelingen der KNAW, afd. Letterkunde, Nieuwe Reeks, deel 31, Nr. 3*. Amsterdam: Noord-Hollandse Uitgevers Maatschappij.

Kommer, M.M. (2009). Personeel. In E.R. Muller & P.C. Vegter (Eds.), *Detentie, gevangen in Nederland* (pp. 355-372). 2e druk, Alphen aan den Rijn: Kluwer.

Kroner, D.G., & Loza, W. (2001). Evidence for the efficacy of self-report in predicting non-violent and violent criminal recidivism. *Journal of Interpersonal Violence, 16*(2), 168-177.

Kravchuk, R.S. & Schack, R.W. (1996). Designing effective performance-measurement systems under the Government Performance and Results Act of 1993. *Public Administration Review*, *56*(4), 348-358.

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.

Kunst, M., Schweizer, S., Bogaerts, S. & Van der Knaap, L. (2008). *Onderlinge agressie en geweld, posttraumatische stress en arbeidsverzuim in penitentiaire inrichtingen*. Den Haag: WODC/Intervict.

Kusek, J.Z., & Rist, R.C. (2004). *Ten Steps to a Results Based Monitoring and Evaluation System: A Handbook for Development Practitioners*. World Bank.

L

Laird, N.M. & Louis, T.A. (1989). Empirical Bayes ranking methods. *Journal of Educational Statistics*, *14*(1), 29-46.

Lambert, E.G., Edwards, C., Camp, S.D. & Saylor, W.G. (2005). Here today, gone tomorrow, back again the next day: Antecedents of correctional absenteeism. *Journal of Criminal Justice*, *33*(2), 165-175.

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.

LaPiere, R.T. (1934). Attitudes vs. actions. *Social Forces*, *13*(2), 230-237.
Leeuw, F.L. (2011). On the effects and perverse effects of performance audits. In J. Lonsdale, P.A. Wilkins & T. Ling (Eds.),*Performance auditing; contributing to accountability in democratic government* (pp. 231-247). Cheltenham: Edward Elgar Publishers.
Leyland, A.H. & Groenewegen, P.P. (2003). Multilevel modelling and public health policy. *Scandinavian Journal of Public Health*, *31*, 267-274.
Liebling, A. (2000). Prison officers, policing and the use of discretion. *Theoretical Criminology*, *4*(3), 333-357.
Liebling, A., & Arnold, H. (2004). *Prisons and their moral performance. A study of values, quality, and prison life*. Oxford University Press: Oxford.
Liebling, A., Durie, L., Stiles, A., & Tait, S. (2005). Revisiting prison suicide: the role of fairness and distress. In A. Liebling, & S. Maruna (Eds.), *The* effects *of imprisonment* (pp. 209-231). Willan Publishing: Uffculme.
Light, P.C. (2006). The tides of reform revisited: Patterns in making government work, 1945-2002. *Public Administration Review*, *66*(1), 6-19.
Lindquist, C.H., & Lindquist, C.A. (1997). Gender differences in distress: mental health consequences of environmental stress among jail inmates. *Behavioral Sciences & the Law*, *15*(4), 503-523.
Lippke, R.L. (2007). *Rethinking Imprisonment*. New York: Oxford University Press.
Logan, C. H. (1992). Well kept: Comparing quality of confinement in private and public prisons. *Journal of Criminal Law and Criminology*, *83*(3), 577-613.
Lonsdale, J. (2011). Introduction. In J. Lonsdale, P.A. Wilkins & T. Ling (Eds.), *Performance auditing; contributing to accountability in democratic government* (pp. 1-22). Cheltenham: Edward Elgar Publishers.
Lonsdale, J. Ling, T. & Wilkins, P. (2011). Conclusions: performance audit - an effective force in difficult times?. In J. Lonsdale, P.A. Wilkins & T. Ling (Eds.), *Performance auditing; contributing to accountability in democratic government* (pp. 311-336). Cheltenham: Edward Elgar Publishers.

M

March, J.G. & Sutton, R.I. (1997). Crossroads - Organisational performance as a dependent variable. *Organization Science*, *8*, 698-706.

Mathison, S. (1988). Why triangulate?. *Educational researcher*, *17*(2), 13-17.
Maxwell, J.A., & Mittapalli, K. (2010). Realism as a stance for mixed methods research. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research* (pp. 145-167). Thousand Oaks: Sage.
McCorkle, R.C., Miethe, T.D., & Drass, K. (1995). The roots of prison violence: a test of the deprivation, management, and "not-so-total" institutional models. *Crime & Delinquency, 41*, 213-32.
Meyer, R.H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*(3), 283-301.
Mills, J.F., Loza, W. & Kroner, D.G. (2003). Predictive validity despite social desirability: evidence for the robustness of self-report among offenders. *Criminal behavior and Mental Health, 13*, 140-150.
Molleman, T. (2008). *Psychometric quality of and the links between the detainee survey and the BASAM-DJI*. Cahier 2008-5, Den Haag: WODC.
Molleman, T. (2011a). *Benchmarking in the prison system: A study on the possibilities of comparing and improving performance.* Den Haag: Boom Juridische Uitgevers.
Molleman, T. (2011b). Ongewenste omgangsvormen tussen gevangenispersoneel, stand van zaken 2011. *Cahier 2011-8*. Den Haag: WODC.
Molleman, T. & Leeuw, F.L. (2012). The influence of correctional staff on prison conditions: a multi-level approach to staff and inmate surveys. *European Journal of Criminal Policy and Research, 18*, 217-233.
Molleman, T. & Van den Hurk, A.A. (2012). Een kwestie van evenwichtskunst: Over doelen en taken van het gevangeniswezen. *Delikt & Delinkwent*, *42*(7), 576-590.

N

Nacci, P.L., & Kane T.R. (1984). Sex and sexual aggression in federal prisons. *Federal Probation*, *48*, 46-53.
Newcomer, K.E. & Triplett, T. (2010). Using Surveys. In J.S. Wholey, H.P. Hatry & K.E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 262-297). New York: Jossey-Bass/Wiley.
Normand, S.L.T., Glickman, M.E., & Gatsonis, C.A. (1997). Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association*, *92*(439), 803-814.

Northouse, P.G. (1997). *Leadership, Theory and Practice*, Thousand Oaks: Sage.
Nyhan, R.C. & Martin, L.L. (1999). Comparative performance measurement: A primer on data envelopment analysis. *Public Productivity & Management*, *22*(3), 348-364.

O

Osborne, D. & Gaebler, T. (1992). *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Public Sector*. Plume.

P

Paterline, A.P., & Petersen, D.M. (1999). Structural and social psychological determinants of prisonization. *Journal of Criminal Justice*, *27*(5), 427-41.
Paulhus, D.L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*, 598-609.

Pawson, R. (2013). *The science of evaluation: a realist manifesto*. Sage.

Pawson, R. & Tilley, N. (1997). *Realistic evaluation*. Sage.
Peterson-Badali, M. & Koegl, C.J. (2002). Juveniles' experiences of incarceration, the role of correctional staff in peer violence. *Journal of Criminal Justice*, *30*, 41-49.
Pidd, M. (2005). Perversity in public service performance measurement. *International Journal of Productivity and Performance Management*, *54*(5/6), 482-493.
Pidd, M. (2012). *Measuring the Performance of Public Services: Principles and Practice*. Cambridge university Press.
Poister, T.H. (2010a). Performance measurement: Monitoring program outcomes. In J.S. Wholey, H.P., Hatry, H. & K.E. Newcomer (Eds.), *Handbook of Practical Program Evaluation* (pp. 100-124). New York: Jossey-Bass/Wiley.
Poister, T.H. (2010b). *Measuring Performance in Public and Nonprofit Organizations*. John Wiley & Sons.
Pollitt, C., & Bouckaert, G. (2011). *Public management reform: A comparative analysis-new public management, governance, and the Neo-Weberian state*. Oxford University Press.
Pollitt, C. (2013). The logics of performance management. *Evaluation*, *19*(4), 346-363.
Pompe, W.P.J. (1950) *Handboek van het Nederlandse Strafrecht*. Derde druk, Zwolle: Tjeenk Willink.

Popper, K.R. (1959). The logic of scientific discovery. *London: Hutchinson*, *1*.
Propper, C., Sutton, M., Whitnall, C., & Windmeijer, F. (2010). Incentives and targets in hospital care: evidence from a natural experiment. *Journal of Public Economics*, *94*(3), 318-335.
Propper, C. & Wilson, D. (2003). The use and usefulness of performance measures in the public sector. *Oxford review of economic policy*, *19*(2), 250-267.

R

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.
Raudenbush, S.W. & Bryk, A.S. (2002). *Hierarchical Linear Models* (2nd ed.). Thousand Oaks: Sage.
Reisig, M.D. & Mesko, G. (2009). Procedural justice, legitimacy, and prisoner misconduct. *Psychology, Crime & Law, 15*(1), 41-59.
Reisig, M.D., & Lovrich, N.P. (1998). Job attitudes among higher-custody state prison management personnel: a cross-sectional comparative assessment. *Journal of Criminal Justice*, *26*(3), 213-226.
Reisig, M.D. (2002). Administrative control and inmate homicide. *Homicide studies, 6*(1), 84-103.
Roth, J. (1985). Consistency of rule application to inmates in long-term treatment institutions. *Social Science & Medicine,* 20, 247-252.

S

Scharff Smith, P. (2009). Solitary confinement: History, practice, and human rights standards. *Prison Service Journal, 181*, 3-11.
Schaufeli, W.B., & Peeters, M.C. (2000). Job stress and burnout among correctional officers: A literature review. *International Journal of Stress Management*, *7*(1), 19-48.
Selden, S.C., & Sowa, J.E. (2004). Testing a multi-dimensional model of organizational performance: Prospects and problems. *Journal of Public Administration Research and Theory*, *14*(3), 395-416.

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.

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.

Snijders, T.A.B. & Bosker, R.J. (2012). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). London: Sage.

Sparks, R., Hay, W. & Bottoms, A. (1996). *Prisons and the Problem of Order*. Oxford: Clarendon Press.

Spivak, A.L., & Sharp, S.F. (2008). Inmate recidivism as a measure of private prison performance. *Crime & Delinquency*, *54*(3), 482-508.

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.

Steiner, B., & Wooldredge, J. (2008). Inmate versus environmental effect on prison rule violations. *Criminal Justice and Behavior*, *35*(4), 438-456.

Steiner, B. (2009). Assessing static and dynamic influences on inmate violence levels. *Crime & Delinquency*, *55*(1), 134-161.

Stiefel, L., Rubenstein, R. & Schwartz, A.E. (1999). Using Adjusted Performance Measures for Evaluating Resource Use. *Public Budgeting and Finance, 19*(3), 67-87.

Stohr, M.K., Lovrich, N.P., Menke, B.A., & Zupan, L.L. (1994). Staff management in correctional institutions: Comparing DiIulio's “control model” and “employee investment model” outcomes in five jails. *Justice Quarterly*, *11*(3), 471-497.

Swanborn, P. G. (1973). *Variabelen en hun meting*. Boom.

Sykes, G.M. (1958). *The society of captives: a study of a maximum security prison*. Princeton: Princeton University Press.

T

Talbot, C. (2010). *Theories of Performance: Organizational and Service Improvement in the Public Domain*. Oxford University Press.

Teddlie, C. & Tashakkori, A. (2010). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In A. Tashakkori & C, Teddlie (Eds.), *Sage*

handbook of mixed methods in social & behavioral research (pp. 3-50). Thousand Oaks: Sage.
Tehrani, M., & Noubary, R. (2005). A statistical conversion technique: objective and perceptive financial measures of the performance construct. *Organizational Research Methods*, *8*(2), 202-221.
Tewksbury, R., & Mustaine, E.E. (2008). Correctional orientations of prison staff. *The Prison Journal*. *88*(2), 207-33.
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.
Tweede Kamer der Staten Generaal (1998). *Penitentiaire beginselenwet*.

U

Useem. B. & Reisig, M.D. (1999). Collective action in prisons: Protests, disturbances, and riots. *Criminology, 37*(4), 735-760.

V

Van Loocke, E. & Put, V. (2011). The impact of performance audits: a review of the existing evidence. In: J. Lonsdale, P.A. Wilkins & T. Ling, T. (Eds.). *Performance auditing; contributing to accountability in democratic government* (pp. 175-208). Cheltenham: Edward Elgar Publishers.
Van Thiel, S., & Leeuw, F. L. (2002). The performance paradox in the public sector. *Public Performance & Management Review*, *25*(3), 267-281.
Vollaard, B. (2003). *Performance contracts for police forces(No. 31)*. CPB Netherlands Bureau for Economic Policy Analysis.
Vuolo, M. & Kruttschnitt, C. (2008). Prisoners' adjustment, correctional officers, and context: The foreground and background of punishment in late modernity. *Law & Society Review, 42*(2), 307-336.

W

Wallace, S.E. (1971). *Total Institutions (cloth)* (Vol. 20). Transaction Books.
Webb, E.J., Campbell. D.T., Schwartz, R.D. & Sechrest, R. (1966). *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.

Wilson, E.O. (1999). *Consilience: The unity of knowledge (No. 31)*. Random House Digital, Inc..

Woodhouse, G., & Goldstein, H. (1988). Educational performance indicators and LEA league tables. *Oxford Review of Education*, *14*(3), 301-320.

Wright, K.N. (1985). Developing the prison environment inventory. *Journal of Research in Crime and Delinquency*, *22*(3), 257-277.

Wright, K.N. (1991). A study of individual, environmental, and interactive effects in explaining adjustment to prison.. *Justice Quarterly*, *8*(2), 217-242.

Wright, K. N. (2005). Designing a national performance measurement system. *The Prison Journal*, *85*(3), 368-393.

Z

Zwezerijnen, J.J.A. (1972). *Dwang en vertrouwen, een empirisch onderzoek naar de machtsrelatie tussen bewaarders en gedetineerden.* Alphen aan den Rijn: Samson Uitgeverij.

Nederlandse samenvatting (Dutch summary)

Prestatiemeting in het Nederlandse gevangeniswezen. Methodologische handvatten voor het meten van prestaties van organisaties in de publieke sector

Afgelopen dertig jaar zijn in veel Westerse landen (semi-) publieke organisaties op afstand van de centrale overheid geplaatst. Hoewel deze organisaties de ruimte werd geboden hun taken naar eigen inzicht uit te voeren, bleven overheden en bewindslieden vaak verantwoordelijk voor de te leveren prestaties. Die verantwoordelijkheid vormde mede aanleiding de prestaties van (semi-) publieke organisaties te gaan meten. Naast het verantwoorden van prestaties en het volgen van prestaties in de loop van de tijd, dient prestatiemeting nog een ander doel. Door de prestaties van verschillende organisaties te vergelijken kunnen zij aangezet worden om te zoeken naar mogelijkheden om hun prestaties te verbeteren en van elkaar te leren.

Naast deze wenselijke effecten is er steeds meer evidentie dat prestatiemetingen ongewenste neveneffecten teweeg kunnen brengen. Organisaties kunnen bijvoorbeeld hun prestaties verbloemen door creatief om te springen met cijfers om zo de (op afstand) verantwoordelijke overheid tevreden te stellen. Een ander ongewenst neveneffect is de verleiding zich vooral op kortetermijndoelstellingen te richten. In dit boek worden methodologische condities van prestatiemeting centraal gesteld die ongewenste neveneffecten tegengaan en prestatieverbetering en verantwoording stimuleren. Het gaat daarbij om het Nederlandse gevangeniswezen.

De eerste conditie die wordt onderscheiden is het hebben van een duidelijk beeld van de doelen en taken van de organisatie waarbij de prestaties gemeten gaan worden. De tweede conditie is het gebruik van valide en betrouwbare manieren van dataverzameling. Daarbij wordt verondersteld dat het risicovol is om bij prestatiemeting slechts op een enkele methode van dataverzameling te vertrouwen. Meetfouten kunnen

namelijk een vertekend beeld van de prestaties genereren. Het gebruik van meerdere meetmethoden kan dat risico verminderen en tegelijk een meer omvattende kijk op de prestaties geven. In dit boek wordt daarom gezocht naar een strategie om verschillende meetmethoden te combineren. De derde onderscheiden conditie bestaat eruit dat rekening wordt gehouden met factoren die de prestaties beïnvloeden, maar niet beïnvloedbaar zijn door het management van de organisatie. Omdat de mate waarin dergelijke factoren een rol spelen verschilt tussen organisaties, kunnen methodologische en statistische procedures behulpzaam zijn bij het verkrijgen van een vergelijkbaar beeld van de prestaties van meerdere organisaties. De vierde en laatste conditie die hier wordt onderscheiden is dat het management van een organisatie de prestaties die worden gemeten ook daadwerkelijk kan beïnvloeden. Acties van het management zijn idealiter gebaseerd op wetenschappelijke evidentie over de werkzaamheid. Dergelijke kennis kan de leer- en verbeterfunctie van prestatiemeting versterken.

De onderzoeksvragen van dit boek volgen uit bovenstaande condities en hebben –zoals gezegd– betrekking op de context van het Nederlandse gevangeniswezen. De eerste onderzoeksvraag luidt: Wat zijn de taken van Nederlandse penitentiaire inrichtingen, volgend uit de doelen van detentie? De doelstellingen van detentie zijn voor het overgrote deel helder; penitentiaire inrichtingen moeten veilige, humane en resocialiserende omstandigheden scheppen. Echter, als we deze doelen in meer detail bezien, lijken complexe tegenstellingen te bestaan binnen een ogenschijnlijk gedetailleerd kader van regels en wetten die betrekking hebben op de uitvoering van straffen en maatregelen. Een gevolg daarvan kan zijn dat penitentiaire inrichtingen verschillen in de wijze waarop zij hun taken vervullen.

We constateren dat het voor een deel duidelijk is welke prestaties worden verwacht van penitentiaire inrichtingen, met name volgend uit de Penitentiaire Beginselenwet en aanverwante regels. Voorbeelden zijn de normen rond het luchten van gedetineerden (ten minste een uur per dag) en het ontvangen van bezoek (ten minste een uur per week). Wet- en regelgeving zijn voor een ander deel van de doelen van detentie een minder heldere leidraad. Twee penologische principes van detentie in Nederland zijn het beginsel van minimale beperkingen en het beginsel van resocialisatie

die worden genoemd in de Penitentiaire Beginselenwet. Omdat deze beginselen verschillend kunnen worden geïnterpreteerd, is het onduidelijk wat de prestaties inhouden die worden verwacht van penitentiaire inrichtingen. Over minimale beperkingen, bijvoorbeeld, meldt de wet dat een penitentiaire inrichting alleen beperkingen aan een gedetineerde mag opleggen wanneer de orde en veiligheid van de inrichting bedreigd worden, of wanneer 'de doelen van detentie' in het geding zijn. Deze doelen van detentie worden echter niet gespecificeerd. Bovendien is het verband tussen de doorgaans in Nederland opgelegde beperkingen en het handhaven van de orde en veiligheid in de inrichting niet vanzelfsprekend. We concluderen dat niet in alle gevallen helder is wat de taken van Nederlandse penitentiaire inrichtingen zijn waardoor onduidelijkheid kan ontstaan over welke prestaties van hen worden verwacht.

De tweede onderzoeksvraag die wordt behandeld in dit boek is de vraag of Nederlandse penitentiaire inrichtingen opereren onder dezelfde contextuele omstandigheden (zoals celcapaciteit, gebouw en regime); en als dat niet zo is, hoe met deze verschillen kan worden omgegaan als we de prestaties van de inrichtingen met elkaar willen vergelijken. Prestatiemetingen kunnen verschillen laten zien in de mate waarin organisaties hun doelen bereiken. Deze organisaties kunnen echter met verschillende contextuele omstandigheden te maken hebben die van invloed zijn op de prestaties. Een vergelijking op basis van ruwe cijfers garandeert daarom niet een accurate Prestatiemeting. Statistische methoden kunnen bijdragen aan het corrigeren van prestatiecijfers om zo de invloed van factoren die buiten de invloedsfeer van het management liggen, in te perken. Als dat niet zou gebeuren, bestaat de kans dat bij een vergelijking van prestaties de best presterende organisatie niet gevonden wordt. Als men vervolgens *good practices* wil uitwisselen tussen organisaties, worden die mogelijk van de verkeerde organisatie 'afgekeken'; namelijk van een organisatie die zijn goede prestatiecijfers dankt aan zijn relatief gunstige contextuele omstandigheden en niet aan zijn effectieve beleid of maatregelen. Een dergelijke situatie staat mogelijk het verbeteren van prestaties in de weg. In sommige (semi-) publieke organisaties, zoals ziekenhuizen, gevangenissen en scholen, is aangetoond dat het corrigeren van prestatiecijfers tot een andere ranking leidt dan de ranking op basis van de ruwe, ongecorrigeerde cijfers. Hoewel de toegevoegde waarde van het statistisch corrigeren

lijkt vastgesteld, wordt het nog maar weinig toegepast. In dit boek stellen we dat een systematisch stappenplan nodig is om tot vergelijkbare prestatiecijfers te komen, waarin de correctietechnieken een plaats hebben. De stappen worden geïllustreerd met twee prestatiemetingen uit het Nederlandse gevangeniswezen.

Ten eerste moet de betrouwbaarheid, validiteit en variabiliteit van de prestatiemeting vast komen te staan. Ten tweede stelt het plan dat beschreven moet worden welke factoren de prestatiescores beïnvloeden. Deze factoren kunnen in drie categorieën vallen, te weten restrictieve factoren (factoren die *buiten* de invloedssfeer van het management vallen), beïnvloedbare factoren (factoren die *binnen* de invloedssfeer van het management vallen) en random variatie (de storingsterm in een stochastisch model). Om te bepalen welke factoren van invloed zijn op een prestatiemeting en in welke categorie de factoren moeten worden geplaatst, wordt een expertraadpleging gebruikt. Ten derde stelt het plan dat de prestatiecijfers moeten worden gecorrigeerd voor restrictieve factoren. Dat kan door de samenhang tussen de prestatiemeting en de factoren te bepalen met een regressiemodel. Met zo'n model wordt vervolgens een voorspelde score per organisatie (hier een penitentiaire inrichting) geschat. Het verschil tussen deze voorspelde score en de gerealiseerde score van een organisatie wordt uitgedrukt in een residu, dat kan worden geïnterpreteerd als dat deel van de score dat kan worden beïnvloed door het management (oftewel: de prestatie). Het stappenplan schrijft tot slot voor alle organisaties te ordenen op hun residu-score wanneer een ranking wordt gemaakt. Als deze stappen worden gevolgd, is de vergelijkbaarheid van organisaties toegenomen en is accuratere en eerlijkere prestatievergelijking mogelijk.

De derde onderzoeksvraag gaat over de veelzijdige aard van thema's waarop (semi-) publieke organisaties moeten presteren. Belanghebbenden kunnen verschillen in hun opvatting over wat de prestaties die deze organisaties moeten leveren exact inhouden en hoe deze te meten. Daardoor is het niet altijd eenvoudig de prestaties van (semi-) publieke organisaties te operationaliseren en meetbaar te maken. Een vaak genoemde oplossing is het gebruik van meerdere meetmethoden zodat de kans afneemt dat uitsluitend op een enkele meetmethode met ernstige meetfouten wordt vertrouwd. Iedere meetmethode kan beperkingen en sterke eigenschappen hebben. Een strategie is

nodig om de prestaties van (semi-) publieke organisaties omvattend te meten, waarbij rekening wordt gehouden met de beperkingen van verschillende meetmethoden. Het gebruik van meerdere meetmethoden zal meetfouten niet voorkomen; maar tegenover de beperkingen van de ene methode kunnen sterke eigenschappen van een andere methode worden gezet. Om de gebruiksvriendelijkheid van informatie over prestaties te bevorderen, zouden verschillende metingen kunnen worden samengenomen in een compositiescore.

In de dit boek wordt het kritisch realisme gebruikt als een epistemologische basis voor het meten van prestaties in de (semi-) publieke sector. De benadering houdt rekening met het fenomeen waarop gepresteerd dient te worden, met onderliggende mechanismen en met contextuele factoren. We veronderstellen dat wanneer meerdere metingen (voortkomend uit verschillende meetmethoden) die betrekking hebben op eenzelfde prestatiethema convergeren, er een bepaald deel van de prestatie adequaat is gemeten. Indien metingen een divergent beeld opleveren, voorziet dit boek ook in een leidraad. Als de betrouwbaarheid en validiteit van twee divergerende metingen vastgesteld zijn, kan een van de twee metingen mogelijk beter passen in een ander prestatiethema dan vooraf werd verondersteld. Om een prestatiethema omvattend te meten, zoeken we naar een bepaalde mate van convergentie, namelijk waarbij metingen statistisch significant samenhangen en ook gecorreleerd zijn in de veronderstelde richting (positief of negatief). De correlaties zouden niet te hoog moeten zijn omdat dan het risico bestaat dat een bepaald aspect van een prestatiethema teveel nadruk krijgt in de uiteindelijke compositiescore. Daarom zouden veronderstelde factoren binnen een prestatiethema – verwijzend naar het fenomeen, het mechanisme en context – gebaseerd moeten zijn op verschillende meetmethoden en een significante maar niet sterker dan een middelgrote samenhang (in de veronderstelde richting) moeten hebben.

Met een prestatiethema uit het Nederlandse gevangeniswezen is gekeken of de hierboven beschreven meetstrategie een leidraad kan vormen voor prestatiemeting in de (semi-) publieke sector. De bevindingen ondersteunen de veronderstelde compositie van het prestatiethema aangaande de *veiligheid van inrichtingspersoneel*. Hoewel veel van de veronderstelde factoren een significante en een tot middelgrote samenhang vertoonden, bleken enkele factoren geen verband met de andere factoren te hebben. Deze bevinding is reden om het samenstel van veronderstelde factoren (over fenomeen,

mechanisme en context) die behoren tot het prestatiethema te heroverwegen in toekomstig onderzoek. Hoe dan ook, bouwstenen zijn gevonden om een prestatiethema 'veiligheid van inrichtingspersoneel' omvattend te meten met meerdere meetmethoden. Deze bevinding ondersteunt de veronderstelde meetstrategie en om die reden wordt geconcludeerd dat we een benadering hebben gevonden die helpt bij het omvattend meten van prestatiethema's in de (semi-) publieke sector.

De vierde onderzoeksvraag gaat over de conditie dat de scores van prestatiemetingen samen moeten hangen met factoren die voor het management van een (semi-) publieke organisaties beïnvloedbaar zijn. Kennis van die samenhang kan van waarde zijn wanneer een organisatie tracht haar prestaties te verbeteren. In dit boek richten we ons op enkele centrale organisatiedoelstellingen van het Nederlandse gevangeniswezen die handelen over de detentieomstandigheden.

Een hypothetisch model is ontwikkeld gebaseerd op de import- en deprivatietheorieën die reacties van gedetineerden op detentieomstandigheden trachten te verklaren. De importtheorie zoekt de oorzaken van reacties van gedetineerden op detentieomstandigheden (zoals percepties en gedrag) in de geschiedenis en eigenschappen van de gedetineerden zelf. Factoren zoals ervaringen in de jeugd van gedetineerden, delictgeschiedenis en persoonlijkheid (-stoornis) worden als verklaringen gezien van de percepties en gedragingen van gedetineerden. De deprivatietheorie brengt de percepties en gedragingen van gedetineerden in relatie met ontevredenheid over hun situatie in detentie. Het ontberen van vrijheid, veiligheid, autonomie en bepaalde voorzieningen zijn daar voorbeelden van. Die deprivaties kunnen bijvoorbeeld worden ingeperkt doordat een penitentiaire inrichting voorziet in positieve contacten tussen gedetineerden en personeel en een eerlijke toepassing van de huisregels.

Om het hypothetisch model te toetsen zijn de resultaten gebruikt van vragenlijstonderzoek onder personeel en gedetineerden. Gebleken is dat op afdelingen waar het personeel zegt een hulpvaardige en steunende bejegeningsstijl te hanteren, gedetineerden hun detentieomstandigheden positiever ervaren. Voorts vinden gedetineerden de regels duidelijker wanneer het personeel een relatief sterke oriëntatie heeft op de huisregels en de structuur van het dagprogramma. Dergelijke factoren – die

binnen de invloedsfeer van het inrichtingsmanagement vallen – zijn relevant gebleken bij de verklaring van gepercipieerde detentieomstandigheden, waarbij gecontroleerd is voor eigenschappen van gedetineerden (zoals etniciteit en leeftijd). De bevindingen suggereren dat import- en deprivatiefactoren complementair zijn bij de verklaring van detentieomstandigheden. De conclusie lijkt gerechtvaardigd dat het management van een inrichting invloed kan hebben op de scores van prestatiemetingen, in dit geval gepercipieerde detentieomstandigheden.

Tot slot is in dit boek gekeken naar de verwachtingen van acht betrokkenen bij prestatiemeting in het Nederlandse gevangeniswezen onder wie directeuren, inspecteurs en experts op het gebied van planning en control. In de interviews vertelden zij de resultaten in deze studie te onderschrijven en positieve verwachtingen te hebben bij toepassing van deze methoden. Zij denken echter dat de methodologische condities van prestatiemeting niet voldoende zijn om verantwoording en prestatieverbetering in het Nederlandse gevangeniswezen te garanderen. Directeuren zouden bij prestatiemetingen bijvoorbeeld niet op het veroordelen of sanctioneren van mindere prestaties uit moeten zijn. Leren van andere organisaties en het verbeteren van de eigen prestaties zijn de belangrijkste doelen van prestatiemeting, aldus de geïnterviewde personen.

We kunnen concluderen dat in dit boek manieren en mogelijkheden zijn ontwikkeld en methodologische oplossingen zijn gevonden ten behoeve van een accurate prestatiemeting in het Nederlandse gevangeniswezen. Hoewel het Nederlandse gevangeniswezen een publieke organisatie is met zijn specifieke eigenheden (zoals organisatiestructuur en -cultuur), is de organisatie onderdeel van een grotere familie van *totale instituties* die complexe doelen nastreven. Verwante (semi-) publieke organisaties zouden daarom kunnen profiteren van de gepresenteerde bevindingen en ontwikkelde methoden. Zulke organisaties zouden kunnen investeren in analyses om hun organisatiemissie en prestatiemetingen aan te laten sluiten, valide en betrouwbare meetmethoden te vinden en een strategie te ontwikkelen om hun prestatiethema's omvattend meetbaar te maken.

In verschillende (semi-)publieke organisaties – in Nederland en daarbuiten – werkt men met instrumenten voor prestatiemeting die aan een of enkele van de genoemde condities voldoen. Echter, we veronderstellen dat als de condities die in dit boek worden genoemd *tegelijkertijd* worden toegepast, de prestatiemeting in sterkere mate aanleiding zal geven tot verbetering en verantwoording van de prestaties. De inzichten in dit boek lijken daarom breder toepasbaar dan in het Nederlandse gevangeniswezen alleen.

Acknowledgements

On the 13th of March 2008 Gerry Gaes and Bo Saylor showed me their performance measurement system at the headquarters of the U.S. Federal Bureau of Prisons in Washington D.C.. Just before that visit I read their book on prison performance. Those experiences made me think about opportunities to improve the performance system in the Dutch prison system. Happily, the Dutch Custodial Institutions Agency (DJI) and the director-general of Prevention and Sanctions (both part of the Ministry of Security and Justice) recognized the importance of proper performance measurement and the thorough developmental work that was needed to bring an effective performance measurement instrument into practice. Therefore, next to Gerry and Bo, a word of thanks is directed to the former head of DJI, Peter van der Sande and the former director-general Dineke ten Hoorn-Boer who supported this project and wanted to ensure the developments for the long term.

In daily contacts during this study I benefitted greatly from conversations with all kind of officials within and outside the Dutch Ministry of Security and Justice. Among them are headquarter officials, business controllers, prison managers, ground staff, colleagues at WODC, scientists (all over the world), inmates, and labor unionists. These people are with too many to give them all the floor here but I am anyhow grateful to you for inspiring me and keeping me focused. However, one person I need to mention in particular and that is Ron Scherf; small but inventive, persistent but comical. Thanks for all those years, from mentorship to friendship.

The guidance I had from professor Frans Leeuw and professor Peter van der Heijden was crucial for this dissertation. Together, we established this PhD-plan in May 2012 to bring different ideas and research directions together. Peter, thanks for your constructive comments and friendly contacts. Frans, our meetings were always inspiring and exciting. Although you can be tough, in light of what I have learned and where I come from, this was more than justified.

Furthermore, I like to thank the members of the manuscript committee of this book, namely professor Kees van den Bos (chair), professor Paul Nieuwbeerta, professor Joop Hox, professor Henk van de Bunt and professor Hans de Groot.

In conclusion, many thanks to my family and friends, Ron and Janneke for being paranimf, my three great brothers, their girlfriends and of course my beloved parents that are always supportive to me. And last but not least Eefje, whose modesty, cheerfulness and beauty makes every day a great day for me.

Amsterdam, May 2014

Toon Molleman

Curriculum vitae

Toon Molleman was born in the Dutch city of Nijmegen but raised in Zieuwent, situated in the heart of the beautiful Achterhoek region in the east of The Netherlands. After obtaining a Master of Science degree in Public Administration at the Free University in Amsterdam, he was employed at the headquarters of the Dutch Custodial Institutions Agency. Since 2007 he is researcher at the Research and Documentation Centre of the Dutch Ministry of Security and Justice. His work mainly focusses on performance measurement, prison conditions, immigrant detention, double bunking in prisons, goals of imprisonment and prison staff issues.

Recent publications

Molleman, T., & Van Ginneken, E. F. (2014). A Multilevel Analysis of the Relationship Between Cell Sharing, Staff–Prisoner Relationships, and Prisoners' Perceptions of Prison Quality. *International journal of offender therapy and comparative criminology*, 0306624X14525912.

Molleman, T., & Van der Broek, T. C. (2014). Understanding the links between perceived prison conditions and prison staff. *International Journal of Law, Crime and Justice*, *42*(1), 33-53.

Beijersbergen, K. A., Dirkzwager, A. J., Molleman, T., van der Laan, P. H., & Nieuwbeerta, P. (2013). Procedural Justice in Prison: The Importance of Staff Characteristics. *International journal of offender therapy and comparative criminology*, 0306624X13512767.

Molleman, T., & Van der Heijden, P. G.M. (2013). A Method to Deal with Dissimilar Circumstances of Public Organizations in Performance Comparisons: Evidence from Dutch Prisons. *Public Administration Research*, *2*(2), p. 1-17.

Alphen, B. van, Molleman, T., Leerkes, A., Van Hoek, J. (2013). *Van bejegening tot vertrek; Een onderzoek naar de werking van vreemdelingenbewaring*, Den Haag: Boom Juridische Uitgevers: Den Haag

Molleman, T. & Van Ginniken, E.F. (2013). We moeten samen het dansje afmaken; PIW'ers aan het woord over belastende en motiverende aspecten van hun werk. *Justitiële Verkenningen*, 39(3), p. 28-40.

Molleman, T. & Leeuw, F.L. (2012). The influence of correctional staff on prison conditions: a multi-level approach to staff and inmate surveys. *European Journal of Criminal Policy and Research,* 18: 217-233.

Molleman, T., Leeuw, F.L. & Bogaerts, S. (2012). De relatie tussen de bejegeningstijl van gevangenispersoneel en de detentieomstandigheden van gedetineerden. Afl. 5, *Sancties*, p. 242-249.

Molleman, T., & Van der Broek, T. C. (2012). *De relatie tussen arbeidssituatie, vakamanschap en detentiebeleving.* p. 65-74 (Hoofdstuk 8). In: Gedetineerd in Nederland 2011, Henneken-Hordijk, I. & Gemmert, N. van. Den Haag: DJI.

Van der Broek, T. C. & Molleman, T. (2012). Tevredenheid van gevangenispersoneel 2011, een verdieping van personele en inrichtingspecifieke kenmerken. *Reeks WODC-cahier.* 2012-2.

Van der Broek, T. C. & Molleman, T. (2012). Personeel in de vreemdelingenbewaring: de arbeidssituatie, agressie en geweld, *Reeks WODC-cahier.* 2012-7.

Molleman, T. & Van den Hurk, A.A. (2012). Een kwestie van evenwichtskunst: Over doelen en taken van het gevangeniswezen. *Delikt & Delinkwent,* 42(7): 576-590.

Leerkes, A. & Molleman, T. (2011). Alternatieve vreemdelingendetentie. *Openbaar Bestuur*, feb., pp. 2-6

Molleman, T. (2011). *Benchmarking in het gevangeniswezen: Een onderzoek naar de mogelijkheden van het vergelijken en verbeteren van prestaties*. Boom Juridische Uitgevers.

Molleman, T. (2011). Ongewenste omgangsvormen tussen gevangenispersoneel, stand van zaken 2011. *Cahier 2011-8*. Den Haag: WODC.

Appendix

This appendix reports on the construction of two survey scales used in this book.

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.

Collegial support (Cronbach's α is 0.83). Items are 5-point Likert scales, ranging from 'totally disagree' to 'totally agree'.

1. My colleagues help me get the work done
2. My colleagues take a personal interest in me
3. I feel at home in this organization
4. My colleagues and I cooperate well
5. My colleagues call me to account when something goes wrong
6. My colleagues are good at their job