University of Strathclyde
Business School

Department of Management Science

OR MODELLING FOR PUBLIC SECTOR PERFORMANCE MEASUREMENT

Robert Bauke van der Meer

Thesis submitted in
partial fulfilment of the requirements
of the University of Strathclyde
for the degree of Doctor of Philosophy

2008

DECLARATION OF AUTHOR'S RIGHTS

SIGNED STATEMENT

This statement is intended clearly to define the extent of my contribution to previously published work for which I have been jointly responsible.

Four papers written by me, together with a variety of co-authors, have been submitted together with the thesis. Three of these papers discuss the application of a variety of OR models relating to the performance of the Maritime and Coastguard Agency (MCA) in coordinating Search And Rescue (SAR) services around the UK coast. The three papers discuss the application of Logistic Regression Analysis (LRA), Data Envelopment Analysis (DEA) and Bayesian Belief Networks (BBN's), respectively, and are thus labelled:

- MCA-LRA
- MCA-DEA
- MCA-BBN

The fourth paper discusses the application of Discrete Event Simulation (DES) to model the performance of a particular Musculo-Skeletal Unit (MSU) operating across two Glasgow hospitals that are both part of NHS Scotland. This paper is thus labelled:

- MSU-DES

I shall now explain my own specific contribution to each of these four papers.

MCA-LRA paper: "Using regression analysis to model the performance of UK Coastguard centres", published in the Journal of the Operational Research Society, Volume 56, Number 6, pages 630-641, 2005.

The authors listed for the MCA-LRA paper are R. B. van der Meer (corresponding author), J. Quigley and J. E. Storbeck. The initial draft, all subsequent drafts and the final version of the paper were written by me. Although all three of us debated how the models might be constructed, I developed the Binary Logistic Regression (BLR) and Stochastic Frontier Analysis (SFA) models that form the heart of the paper. I carried out all calculations and performed all statistical tests for the models included in the paper – although I discussed these extensively with John Quigley, in particular, at various stages of the analysis. I also composed the introductory and concluding Sections of the paper.

As members of the research team, both John Quigley and James Storbeck contributed substantially to our initial debates about the theoretical form and empirical scope of the research. As the research got under way, John Quigley contributed mainly to the development and analysis of the BLR model – in particular, by discussing with me the nature of the disturbance term in the BLR model and suggesting appropriate statistical tests such as the Hosmer-Lemeshow goodness-of-fit test. (I took sole responsibility for developing the Stochastic Frontier Analysis model.) James Storbeck's principal contribution was to the MCA-DEA paper (see below).

MCA-DEA paper: "Using data envelopment analysis to model the performance of UK Coastguard centres", published in the Journal of the Operational Research Society, Volume 56, Number 8, pages 889-901, 2005.

The authors listed for the MCA-DEA paper are R. B. van der Meer (corresponding author), J. Quigley and J. E. Storbeck. The initial draft and the final version of the paper were written by me. As for intermediate drafts, I was solely responsible for all Sections, except the Section on "DEA analysis of the data", for which various drafts were written jointly by me and James Storbeck; that is, I wrote the initial draft of this Section, which was extended by James Storbeck and put into a final form by me. The construction of the DEA model was extensively discussed between James Storbeck (who is an expert on DEA modelling) and me. However, I carried out the initial calculations, which were subsequently checked and extended by James Storbeck, and then checked and finalised by me.

James Storbeck contributed substantially to the development and analysis of the DEA models – in particular, by suggesting that the length-of-coastline (LC) variable should be used as a non-discretionary output in DEA Model 2 and also by suggesting that, in each of the two DEA models, overall technical efficiency should be decomposed into the 'within-year' and 'between-years' technical efficiencies. John Quigley's principal contribution was to the MCA-LRA paper (see above) and, in particular, the MCA-BBN paper (see below).

MCA-BBN paper: "Modelling the reliability of search and rescue operations with Bayesian belief networks", published in Reliability Engineering & System Safety, Volume 93, Number 7, pages 940-949, 2008.

The authors listed for the MCA-BBN paper are L. Norrington, J. Quigley (corresponding author), A. Russell and R. B. van der Meer. John Quigley was responsible for the final version of the paper. I wrote the initial draft of Sections 1 (Introduction) and Section 2.1 (Background – UK maritime rescue). Lisa Norrington wrote the initial drafts of Sections 2.2 (Background – summary of literature on maritime risk research) and 4 (Elicitation of BBN); and John Quigley wrote the initial draft of Section 3 (Secondary data analysis) and Section 5 (Summary and future work).

The methodology to develop the BBN was suggested by John Quigley (who is an expert on BBN modelling). I mainly contributed my knowledge and understanding of Search And Rescue operations, which – together with the contingency tables constructed by John Quigley – formed the initial basis for the development of the BBN. The subsequent fieldwork (principally interviews with relevant experts) underlying the development of the BBN was carried out initially by Ashley Russell (as part of her MRes dissertation) and completed by Lisa Norrington (as part of her MSc OR dissertation). Both Ashley Russell's and Lisa Norrington's projects were jointly supervised by John Quigley and me (with me taking primary responsibility for supervising Ashley Russell's project and John Quigley doing the same for Lisa Norrington's project). Together, John Quigley and I checked the empirical results obtained by Ashley Russell and Lisa Norrington and decided on the structure of the paper.

MSU-DES paper: "Using OR to support the development of an integrated musculo-skeletal service", published in the Journal of the Operational Research Society, Volume 56, Number 2, pages 162-172, 2005.

The authors listed for the MSU-DES paper are R. B. van der Meer (corresponding author), L. A. Rymaszewski, H. Findlay and J. Curran. The initial draft, all subsequent drafts and the final version of the paper were written by me. The paper discusses the results of a series of OR projects carried out by Management Science students between 1999 and 2003. I designed all of these interventions and, therefore, the student projects were supervised by me, with Lech Rymaszewski (NHS consultant surgeon) and Helen Findlay (NHS senior nurse) as clients. The final project included in this series was carried out between September 2002 and April 2003 by Jonathan Curran as part of his Honours dissertation. Although Jonathan Curran did not contribute any text to the paper (either directly or indirectly), Figures 1 – 6 were taken from his client report. After I had completed the final draft of the paper, I discussed it with Lech Rymaszewski and Helen Findlay – Jonathan Curran having left the University after completing his degree – but no changes (either minor or major) were made at that point.

Signed: **Date: 13 March 2008**

ACKNOWLEDGEMENTS

This thesis is dedicated to the memory of my parents Bauke van der Meer and Nel van der Meer-Bos.

I would very much like to thank my partner Mairi and my daughters Martiena and Carola for their love and support. I would also like to thank all my other relatives and friends in Holland (in particular, Peter, Tineke, Arjen, Maarten and Eva) and Scotland for keeping faith in me. I am very grateful to my colleagues in Management Science for making the Department a fun place to work in, with special thanks to my research collaborators Dr John Quigley, Professor James Storbeck (now of the University of Texas), Lisa Norrington, Ashley Russell, Mr Lech Rymaszewski (NHS Scotland), Helen Findlay (NHS Scotland) and Jonathan Curran. I would like to express my sincere gratitude to Professor Norman Lawrie, who was my first mentor in Management Science, and Professor Tim Bedford and Professor Lesley Walls for their encouragement and good advice.

Last, but not least, a big thank you to the staff of NHS Scotland. Without their skill and dedication I would not have been able to write this thesis.

CONTENTS

LIST OF FIGURES

ABSTRACT

The overall research theme of this thesis is to analyse, from both a theoretical and an empirical perspective, how Operational Research modelling can be used in support of public sector performance measurement. The thesis includes four papers written by Van der Meer together with a variety of co-authors. Three of these papers discuss the application of a variety of OR models – in particular, Logistic Regression Analysis, Stochastic Frontier Analysis, Data Envelopment Analysis and Bayesian Belief Networks – relating to the performance of the Maritime and Coastguard Agency in coordinating Search And Rescue services around the UK coast. The fourth paper discusses the application of Discrete Event Simulation to support performance measurement in a Musculo-Skeletal Unit operating across two Glasgow hospitals that are both part of NHS Scotland.

The thesis investigates five research questions, all of which can be directly traced back to the overall research theme. The evidence base for the investigation consists of the papers referred to above, in conjunction with a wide range of sources from the academic and professional literature. The conclusions may be briefly summarised as follows. First, there are five specific problems and challenges for performance measurement in public service organisations. Second, there are at least two categories of OR modelling approaches that can play an important role in support of public sector performance measurement; namely, models for measuring performance differentials and models for understanding the cause-and-effect relationships driving such performance differentials. Third, while there are advantages to a multimethodology approach to OR modelling, there are also a number of barriers (including paradigm incommensurability). Fourth, the adoption of Critical Realism as a philosophical basis for combining different OR models can resolve problems of paradigm incommensurability. Fifth, OR models that are founded on a realist ontology should be validated according to two specific groups of criteria.

1 INTRODUCTION

1.1 Explanation of the research theme

The overall research theme of this thesis is to analyse, from both a theoretical and an empirical perspective, how Operational Research (OR) modelling can be used in support of public sector performance measurement. The thesis consists of nine chapters (plus a set of references constituting Chapter 10) written by Van der Meer and also four papers written by Van der Meer together with a variety of co-authors that are reproduced in Chapters 11 – 14, respectively. (For the latter, please see the "signed statement" printed above.) All four of these papers deal with Operational Research (OR) models for performance measurement in the public sector; they have been published in recent volumes of reputable, peer-reviewed academic journals.

Three of these papers discuss the application of a variety of OR models relating to the performance of the Maritime and Coastguard Agency (MCA) in coordinating Search And Rescue (SAR) services around the UK coast. The three papers discuss the application of Logistic Regression Analysis (LRA), Data Envelopment Analysis (DEA) and Bayesian Belief Networks (BBN's), respectively, and are thus labelled:

- MCA-LRA (Van der Meer, Quigley & Storbeck, 2005a)
- MCA-DEA (Van der Meer, Quigley & Storbeck, 2005b)
- MCA-BBN (Norrington, Quigley, Russell & Van der Meer, 2008)

The fourth paper discusses the application of Discrete Event Simulation (DES) to support performance measurement in a particular Musculo-Skeletal Unit (MSU) operating across two Glasgow hospitals that are both part of NHS Scotland. This paper is thus labelled:

- MSU-DES (Van der Meer, Rymaszewski, Findlay & Curran, 2005)

Whereas each of these four papers deals with a mix of theoretical and empirical questions relating to a specific public sector context (and can therefore be seen, at least to some extent, as case studies), the nine chapters of this thesis are primarily intended to provide the analytical framework for placing the papers under the umbrella of the overall research theme. To focus our analysis, we shall pose a number of specific research questions next.

1.2 Research questions

The four papers under consideration deal with different examples of OR modelling in support of public sector performance measurement. In the next eight chapters of this thesis we shall, however, seek to explore a number of common questions that can be directly traced back to the overall research theme. The first three of these questions were largely inspired by research issues of immediate concern to most, or all, of the four papers.

All four papers are concerned with performance measurement in public service organisations. Public service organisations are those parts of the public sector that, directly or indirectly, provide various kinds of essential services to members of the public (either the population as a whole or specific eligible groups). They are not profit making and tend to be paid for out of general taxation. The particular nature of public service organisations raises the question of whether they face problems and challenges in relation to performance measurement that are not shared – or shared to a lesser extent – by, say, private-sector business organisations. This forms the subject of the first research question.

Another common feature of the four papers is that they involve the application of OR models in support of performance measurement. However, the particular modelling approaches are different in each case. Therefore, one can ask a second research question about the various ways in which OR modelling can support performance measurement in public service organisations – more specifically, what might be regarded as the purpose(s) of OR modelling, how might particular OR models be selected and developed, and to what extent can more general principles of OR modelling be applied within this overall context?

The third research question is directly inspired by the fact that the three MCA papers apply different modelling approaches the same issue; namely, the performance of the Maritime and Coastguard Agency in coordinating Search And Rescue services. In other words, the three MCA papers may together be taken as a particular example of a multimethodology approach to OR modelling. The question is then what kind of role a multimethodology approach can play within this overall context – more specifically, what are the advantages of such an approach and what are the potential barriers?

A consideration of the potential barriers faced by OR practitioners wanting to try a multimethodology approach leads to the final two research questions to be investigated in the next eight chapters. It will be argued that a multimethodology approach to OR modelling lends added weight to the usual questions about what kind of research philosophy to adopt and what type of model validation criteria to apply within the overall context of performance measurement in public service organisations.

In summary, the five research questions are as follows:

1. What are the particular characteristics of performance measurement in public service organisations?
2. How can OR models be applied in support of performance measurement in public service organisations? To be more precise:
    a. which particular OR models could be used and to what purpose;
    b. and how applicable are more general principles of OR modelling within this overall context?
3. What kind of role can a multimethodology approach to OR modelling play in support of performance measurement in public service organisations? (What are the advantages, and what are the potential barriers?)
4. Given the adoption of a multimethodology approach within this overall context, what kind of research philosophy should be adopted?
5. Given the adoption of a multimethodology approach within this overall context, what type of model validation criteria should be applied?

The investigation of the first three of these research questions will start by examining key points in the academic and professional literature on performance measurement in public service organisations, which is but a subset of the vast, and ever-growing, academic and professional literature on organisational performance measurement in general. The findings from the literature will then form the basis for a thorough review of the MCA and MSU papers in terms of their specific organisational contexts and the range of OR models applied. The last two research questions, on the other hand, will primarily be explored through a detailed discussion of those elements in the (wide-ranging) academic literature on research philosophy and model validation that would appear to be particularly relevant to OR modelling in support of performance measurement in public service organisations.

1.3 Initial discussion of the boundaries to the research

The overall research theme of this thesis is very broad. In principle, it would apply to different kinds of public sector organisations, to different forms of performance measurement and also to different types of OR modelling approaches. But the four papers that form our own contribution to the empirical basis for this thesis are, inevitably, rather more limited in their scope. They deal with two instances of public service organisations rather than the public sector as a whole, and with a limited range of performance measures and OR modelling approaches. Obviously, one should always be very wary about making inductive generalisations from such a small number of examples.

However, the theoretical and empirical evidence from the academic literature cited in this thesis is meant to have a much wider applicability. By carefully reflecting on the empirical findings from the four papers in the light of the overall analytical framework constructed from the relevant academic literature, it is hoped that this thesis can provide convincing answers to the research questions listed above – and that these answers will have a much wider validity than could be inferred from the limited empirical scope of the four papers when seen in isolation from the other parts of the academic literature. We shall return to this issue when we are discussing our overall conclusions in Chapter 9 of this thesis.

1.4 Brief outline of the thesis

Including this introductory chapter, the thesis consists of fourteen chapters in total. Chapters 2 – 5 all follow a similar format. First, key points in the relevant academic and professional literature are highlighted. Next, these points form the basis of a thorough review of the actual findings from the MCA and MSU studies, respectively. The progression from Chapter 2 to Chapters 3, 4 and 5 is guided, at least to some extent, by the first three steps in the OR modelling process suggested by Mingers and others (Mingers & Brocklesby, 1997; Mingers, 2006a):

1. Appreciation of the research situation as experienced by the researchers involved and expressed by any actors in the situation, and prior literature and theories.
2. Analysis of the information from the first phase so as to understand the history that has generated it, and the particular structure of relations and constraints that maintain it.
3. Assessment of the postulated explanation(s) in terms of other predicted effects, alternative possible explanations, and, within action research, consideration of ways in which the situation could be other than it is.
4. Action to bring about changes if necessary or desired, or to report on and disseminate the research results.

Chapter 2 is intended to enable an initial appreciation of the research situation by providing the overall context for performance measurement in the public sector in general, and public service organisations in particular. Chapter 3 is related to the second step in the modelling process by examining the particular characteristics – in terms of organisational problems and challenges – of performance measurement in public service organisations. From Chapter 4 onwards, the attention shifts to the assessment of possible explanations of organisational performance through the use of OR models, which could also suggest ways in which performance might be improved. Chapter 4 discusses what purposes could be served by OR modelling and

Chapter 5 considers which particular OR models could be used to understand cause-and-effect relationships affecting the performance of public service organisations.

Chapter 6 focuses on the advantages of, and potential barriers to, a multimethodology approach to OR modelling and reviews the way in which this approach was used in the MCA studies. From this arise questions about the most appropriate research philosophy to be adopted and the type of model validation criteria to be applied in these kinds of studies – questions which are discussed in Chapter 7. Chapter 8 provides a set of final reflections on the four papers under consideration, including comparisons of the papers to each other and to studies conducted by other authors; while the concluding Chapter 9 discusses the summary findings of this thesis in relation to the research questions listed in Section 1.2, reviews the boundaries of this research and suggests some questions for further research.

Finally, the references for the first nine chapters are listed in Chapter 10; and Chapters 11 – 14 consist of copies of the four previously-published papers under consideration.

2 THE CONTEXTS OF THE OR MODELS

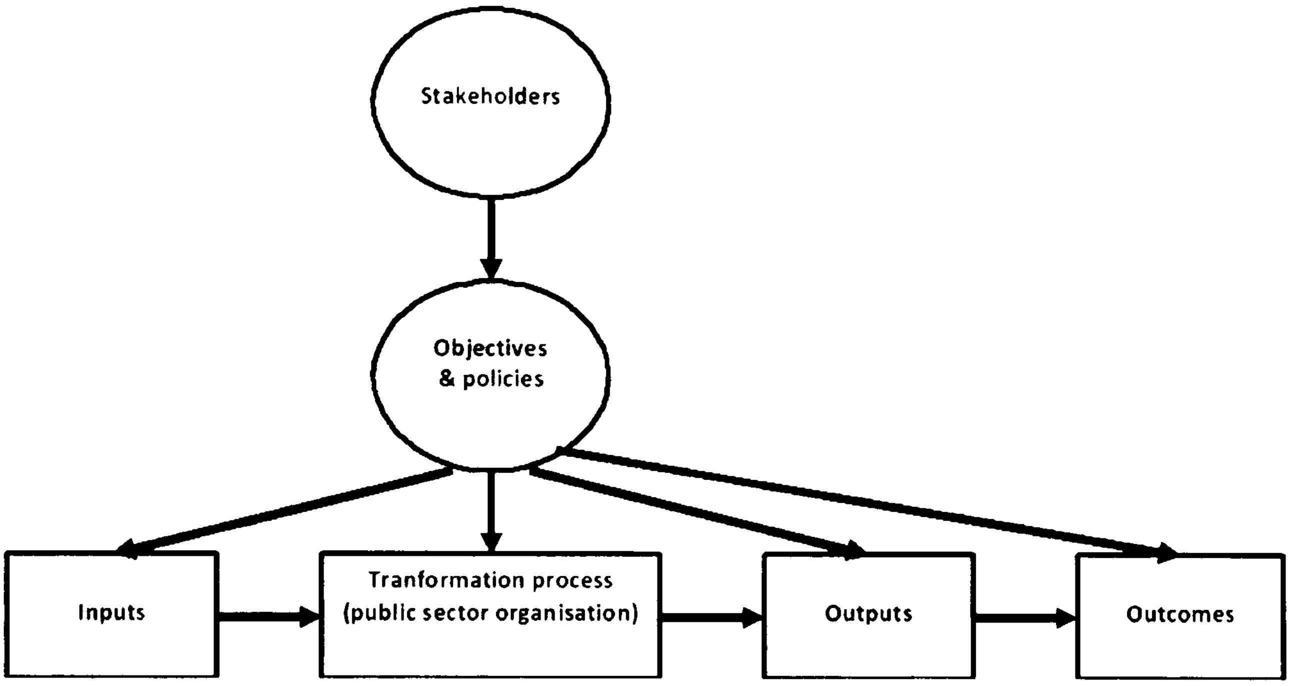
2.1 General introduction to performance measurement in the public sector

As Mingers (2006a) has pointed out, the first phase in any process of OR modelling is to gain an appreciation of the overall research situation. The modelling context for all four of the papers under consideration is performance measurement in the public sector – or, more specifically, in public service organisations. Judging from the large amount of academic literature on this topic (we shall quote a number of these references below), public service organisations appear to constitute one of the parts of the public sector in which problems of performance measurement are felt most acutely.

Performance measurement in public sector organisations is commonly analysed using a relatively simple conceptual model, whereby inputs are transformed into outputs and, eventually, into outcomes. These outcomes are the consequences of the outputs produced in an organisation (Van Peursem *et al.*, 1995), as shown in Figure 1 overleaf.

In the case of public service organisations, the outputs tend to be (intangible) services rather than (tangible) goods. Also, both the outputs and the outcomes may involve co-production with the service users.

Figure 1. A simple transformation model of organisational performance



(Adapted from Johnsen, 2005, Fig. 1, and Jacobs *et al.*, 2006, Fig. 2.1)

There are a range of issues that must be considered when selecting performance measures relating to outputs and outcomes (Anthony & Young, 2002), namely: objective versus subjective measures; quantitative versus non-quantitative measures; performance measured on a continuous scale versus discrete measures; actual versus surrogate measures; and measures primarily related to quantity versus those primarily related to quality. Problems of data availability and measurement errors typically lead to the adoption of performance indicators (PI's) rather than more precise measures of organisational performance.

In general, there are four summary measures of performance: economy, efficiency, effectiveness, and equity (Van Peursem *et al.*, *op. cit.*; Boland & Fowler, 2000; Johnsen 2005). (That is, there are four E's, rather than just three as is often claimed.) Economy is concerned with the cost, number or quality of the inputs. Efficiency measures the ratio of outputs to inputs. Effectiveness measures the extent to which the outputs and, more particularly, the outcomes meet the objectives set for the service (given the inputs consumed). Equity measures the extent to which the outputs and outcomes meet considerations of fairness among service users (e.g. equal opportunity of access). The four E's are typically measured using composite performance indicators.

Note that the definitions of these terms are not always consistent between different authors. Other terms may be also be used, such as productivity. Like efficiency, the concept of productivity relates to the ratio of some (or all) of the outputs to some (or all) of the inputs. More confusingly, in 'Frontier Analysis' – such as Stochastic Frontier Analysis (SFA) and Data Envelopment Analysis (DEA) – the term efficiency tends to be treated as synonymous with cost-effectiveness; that is, a composite performance indicator measuring the relationship between the total value of the outcomes achieved and the total cost of the inputs consumed (cf. Jacobs *et al.*, 2006).

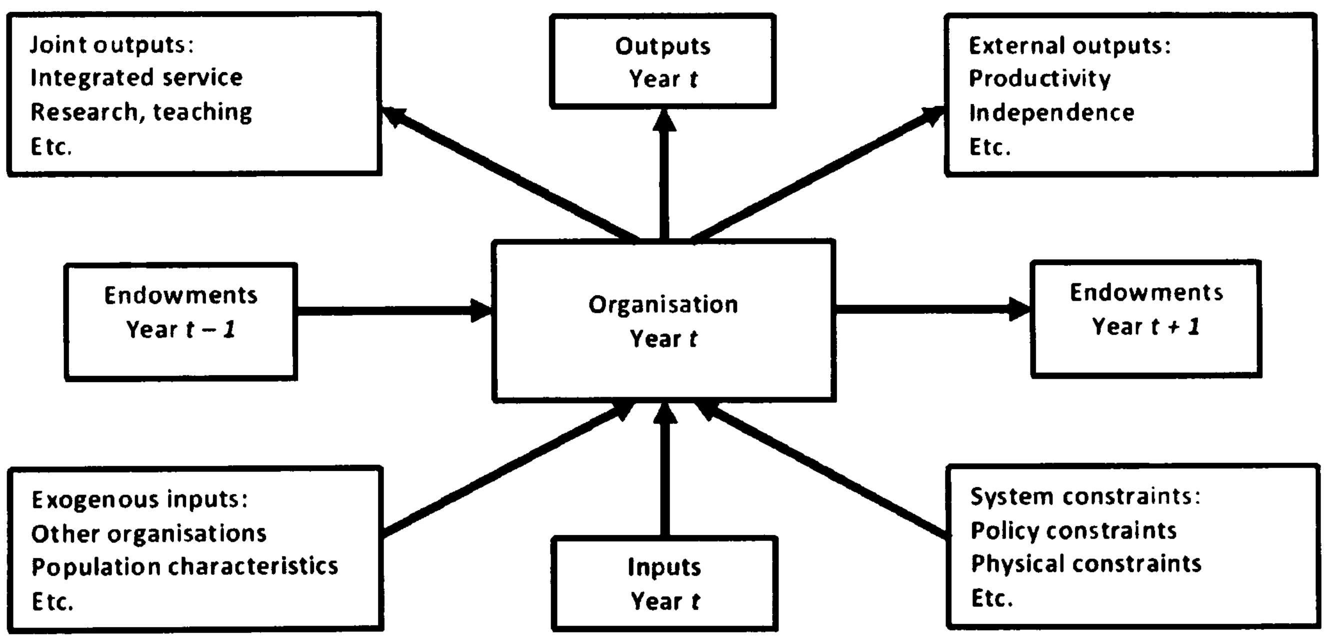
Although this kind of simple performance model can, in principle, also be applied to business organisations in the private sector, the situation is more complicated in the public sector. If the outputs are provided free at the point of delivery, then 'willingness to pay' on the part of service users is not a feasible indicator for the quality of these outputs. Instead of looking at output or outcome measures, one could focus attention on efficiency (or productivity) measures. But as is noted by Williams (1993, pp. 653-654), there is again a problem.

> *"Productivity is certainly related to outcomes, but the relationship is not straightforward. A very productive organisation – that is, one that very efficiently converts resources to outputs – may yet perform poorly if the outputs do not contribute to socially valued objectives. This enigma is at the centre of the difference between the public and private sectors. In the private sector, efficient production of poorly valued outputs is unprofitable. The public sector may not experience this penalty."*

So the outputs and outcomes of public service organisations must be valued somehow. This value cannot be measured by the willingness to pay on the part of the service users but depends, instead, on the subjective valuations of all of the stakeholders involved – broadly speaking, the providers and funders of the service as well as its users. But these different stakeholders do not necessarily use the same yardsticks to arrive at their valuations.

Moreover, the picture presented by the simple performance model of Figure 1 is incomplete. Rather than a simple transformation process, the public service organisation should be seen as the central element in a complex system, as shown in Figure 2 below, which is embedded in an environment characterised by uncertainty and ambiguity (Johnsen, 2005).

Figure 2. A systems view of organisational performance



(Adapted from Jacobs *et al.*, 2006, Fig. 2.5)

Figure 2 shows that both inputs and outputs are more complex than is often assumed and that the organisation's history is important. System inputs should be taken to include exogenous inputs, system constraints and previous investments in the service. System outputs (and outcomes) incorporate joint outputs with other organisations, external outputs and the future resource base for the service. (A detailed explanation of specific organisational problems and challenges in public service organisations is provided in Section 3.1 below.)

Public service performance analysis has three broad aims (Bird *et al.*, 2005, p. 2):

1. to establish 'what works' in promoting stated objectives of the public services;
2. to identify the functional competence of individual practitioners or organisations;
3. public accountability by Ministers for their stewardship of the public services.

The second aim includes examining the scope for improvements in public service provision; this aim has tended to predominate in practice. As we shall discuss below, the first aim is highly important for actually achieving such performance improvements but is widely regarded as problematic.

The third aim relates to the Government's dual role of both monitoring public services and being monitored by performance indicators. With respect to its latter role, the Government's actions require independent scrutiny and well-informed public debate (Bird *et al.*, *op. cit.*; Smith, 2005). Smith (1995b, p. 13) argues that:

> *"... it is unlikely that a consensus will emerge as to how performance in the public sector should be interpreted ... evaluation therefore depends on the objectives of the individual undertaking the analysis. So the proper role of [public sector] performance indicators should be to inform political debate, and more attention should be given to the democratic institutions that allow that debate to take place."*

Section 5.3 below provides an explanation of four different categories of OR modelling approaches. But at this point it is useful to note that, in order to measure public sector performance, OR practitioners have at their disposal a number of Frontier Analysis techniques focusing on composite indicators of efficiency (or cost-effectiveness) (Jacobs *et al.*, 2006). These are Stochastic Frontier Analysis (SFA), which is a parametric method based on a particular form of regression analysis, and Data Envelopment Analysis (DEA), which is a non-parametric method. In the context of our own work, the application of SFA is discussed in the MCA-LRA paper and is compared with the application of DEA in the MCA-DEA paper. We shall examine some issues relating to the practical use of SFA and DEA in later chapters; in particular, Sections 5.5 and 6.2.

However, as we shall argue in Sections 5.1 – 5.3, performance measurement on its own is not enough. Frontier Analysis techniques may help us to uncover significant performance differentials. But these methods can be complemented with other OR modelling approaches – such as Discrete Event Simulation (DES) or Bayesian Belief Nets (BBN's) – that enable us to gain a better understanding of the cause-and-effect relationships driving these performance differentials. In the context of our own work, the development of a BBN is discussed in the MCA-BBN paper and its role in our multimethodology approach is further explored in Sections 5.5 and, more particularly, 6.2. The application of DES, on the other hand, is discussed in the MSU-DES paper and further analysed in Section 5.6.

2.2 The general context for performance measurement in the Maritime and Coastguard Agency

The MCA-LRA, MCA-DEA and MCA-BBN papers are all concerned with the performance of Coastguard coordination centres, operated by the Maritime and Coastguard Agency. As their name implies, the main task of these centres is to coordinate the Search And Rescue (SAR) response to incidents around the UK coast. (The SAR response itself is typically provided by teams of local volunteers.) This is a vital but complex public service, which is provided free to all those involved in such incidents. The research reported in the three MCA papers focused on variations in performance between different coordination centres (each centre being associated with its own Coastguard district) and between successive years.

Human and capital resources constitute the main inputs used by the coordination centres: human resources in the form of various grades of watchkeeping officers and capital resources in the form of buildings and various types of technical equipment (such as advanced communication systems). The measurable outputs provided include the total number of incidents dealt with by each coordination centre in each year, the number of incidents in which assistance was rendered, the number of persons assisted, the number of persons rescued and, as a 'negative' output, the number of lives lost. It is generally agreed – for maritime rescue operations, see, e.g., Skyttner & Fagrell, 1998; for other emergency services such as the ambulance service, see, e.g., Heath & Radcliffe, 2007 – that a key outcome of an emergency service is the proportion of lives saved (of all those involved in life-threatening incidents).

There was some hint in the various government reports produced on this issue in the late 1990's (see the MCA-LRA paper for the precise references) that the Government was particularly concerned with labour productivity as a performance measure (with labour productivity being felt to be too low). However, to be fair to the Government's position, the effectiveness of the service in terms of its ability to produce successful outcomes (i.e. lives saved) was also thought to be very important

– as is clear from Lord Donaldson's report (1999) commissioned by the Government. More generally, it will be clear from this brief description that there are a range of stakeholders involved in this service, each group potentially having a different view on the cause-and-effect relationships underlying the performance of the service. But we shall return to that point in Section 5.5.

It is important to take a systems view of organisational performance. The MCA is a rather traditional organisation (many of its staff used to come from the Royal Navy or the merchant navy) and both its recent history and its intended future state are relevant to a proper understanding of the issue. The performance service of the service is affected by exogenous inputs such as changes in people's leisure activities around the coast and the now widespread availability of mobile phones, and system constraints such as shifts in Government policy affecting funding etc. Also, although the coordination of SAR operations is generally considered to be the main task of the MCA, there are other outputs such as educating members of the public about how to avoid unnecessary risks and how best to act if emergencies do occur. The outputs and outcomes of the Coastguard service are co-produced both with the teams of local volunteers directly providing the SAR response and the service users themselves. Therefore, when performing efficiency analysis based on SFA or DEA, we should try to incorporate as many environmental factors (i.e. factors that affect the performance of the Coastguard service but are not directly under the control of the management of the service) as possible.

Although the Government, who funds the MCA directly out of general taxation, may be mainly concerned with evaluating the functional competence of MCA staff and identifying opportunities for improving the running of the service, the most important aim for the research discussed in our three MCA papers has been to inform public debate – in particular, to gather evidence on whether the Government's actions over the last decade to reorganise key elements of the service have had the positive effects on performance that were intended. We shall discuss the purpose of the various MCA models in some more detail in Chapter 4.

2.3 The general context for performance measurement in the Musculo-Skeletal Unit

The MSU-DES paper is concerned with efforts to improve the performance of the Musculo-Skeletal Unit operating across two Glasgow hospitals that are both part of NHS Scotland. The main task of this Unit is to provide appropriate health care to patients with a range of musculo-skeletal disorders, on both an outpatient and an inpatient basis. This is a complex service, which integrates the activities of orthopaedics and rheumatology with specialist physiotherapy and podiatry. It is provided free to all NHS patients requiring this service. The research reported in this paper focused on developing a simulation model to analyse and, where possible, improve patient waiting times as a key performance indicator for the service.

The main inputs are, as usual in the health service, a wide range of human and capital resources. The latter consist of the hospital buildings and all the necessary physical resources that they contain (beds, operating theatres, etc.). In the present case, however, the key inputs – one might say the bottleneck resources – are the medical and nursing staff involved (hospital consultants, specialist nurses, physiotherapists, podiatrists and unit administrators). The measurable outputs may be grouped into two broad categories: the level of service provided to patients (as indicated by variables such as the number of patients treated over a given time period and the length of patient waiting times) and the efficiency with which the inputs are utilised. The main outcome would be actual improvements in the health status of eligible patients as a (more or less) direct result of providing the service. But as such outcome data were not readily available for this research, attention focused instead on output measures of performance – in particular, as already noted, patient waiting times.

In a public service of this kind, all four of the E's (economy, efficiency, effectiveness and equity) are, in principle, very important. While the main focus of the clients for this research (the senior medical and nursing staff in charge of the Unit) tended to be on considerations of effectiveness and equity, it is probably fair to say that the

(Scottish) Government, as the funding body for the NHS, was concerned at least as much with questions of economy (keeping NHS spending within annual budgets) and efficiency (reducing any wastage of resources). In such situations, the potential trade-off between service level and efficiency or, more generally, the balance between effectiveness and equity versus efficiency and economy becomes a key issue for investigation. Other stakeholders involved will have their own views on how this kind of trade-off should be handled, but we shall return to that point in Section 3.3.

As with the Coastguard service, it is important to take a systems view of organisational performance in the Musculo-Skeletal Unit. The Unit underwent substantial changes during the period under study. It was established at one Glasgow hospital (Stobhill) and was then reorganised and expanded to incorporate another, larger one (Glasgow Royal Infirmary). The performance of the Unit is influenced by exogenous inputs such as the comparatively poor health status of many people in Glasgow's East End, and by system constraints such as shifts in Government policy (for instance, the implementation of devolution for Scotland and, as a separate issue, the increasing emphasis on reducing patient waiting times) affecting the organisation and management of the NHS as a whole. Also, the Unit produces a range of important outputs jointly with other organisations, such as local GP practices (in its ongoing attempts to develop a truly integrated service for patients) and with local Universities (in terms of its teaching and research functions). The outputs and, more particularly, the outcomes of the service are co-produced with the service users; that is, the patients attending the outpatient clinics or using the inpatient facilities provided. Therefore, when evaluating the performance of the Unit we should, wherever possible, try to take such environmental factors into account.

Unlike the Coastguard service, however, the main aim for the research discussed in this paper has not been so much to open up the service more effectively to public accountability, but to help our clients in finding ways to improve the organisation and management of the Unit and thus its performance. In consequence, the research concentrated on performance analysis through simulation modelling rather than overall performance measurement through Frontier Analysis. We shall discuss the purpose of the MSU-DES model in some more detail in Chapter 4.

2.4 Data collection problems in the OR studies relating to the Maritime and Coastguard Agency

As part of the appreciation of the research situation and as a prerequisite for its subsequent analysis, data must be gathered on relevant input, output and (where possible) outcome variables, as well as any important environmental factors. However, performance measurement of public service organisations may be bedevilled by the lack of accurate and reliable data, which may necessitate the use of a range of indicators rather than precise measures to assess performance (Anthony & Young, 2002; Boland & Fowler, 2000; Collier, 2006; Jacobs *et al.*, 2006). Apart from possible disagreements about quantity measures, different stakeholders in a public service may, as noted by Wisniewski & Stewart (2004), want to make their own (subjective) judgements on the quality of important outputs or outcomes – adding to the potential confusion surrounding this issue.

In the MCA studies, the main set of available data consisted of the annual SAR statistics for each Coastguard district. These data form part of official publications by the Government and are, therefore, readily available. (Although that is not true for one of the explanatory variables used in the regression models; namely, the length of coastline monitored by each coordination centre. Those data had to be requested separately from the MCA and, after the closures of some centres in 2000 and 2001, had to be re-estimated by us.)

The annual Coastguard statistics are aggregated from individual incident records kept at MCA headquarters in Southampton, which are not publicly available. In the course of the studies, however, we became aware from conversations with some of the stakeholders involved (MCA watchkeeping officers; the trade union official representing the officers) that these individual incident records are likely to contain various measurement errors. It is one of the responsibilities of watchkeeping officers to record key details during or shortly after each incident. However, in practice, these details do not always seem to be recorded accurately and there may also be differences in interpretation between watchkeeping officers about the precise definitions of important variables. We could only start investigating some of these issues ourselves when we were finally given access to one year's set of individual incident records in the course of the development of a Bayesian Belief Net for the service. (However, there is no real mention in the MCA-BBN paper of this particular element of the research, as it had not yet been completed when that paper was submitted for publication.)

In our analysis, we adopted the 'odds ratio' – that is, the ratio of the number of people rescued to the number of lives lost – as the key outcome indicator for the service. However, because we did not have access to the individual incident records, we were unable to distinguish between the different types of fatal incidents. In particular, we could not distinguish in our analysis between 'genuine' maritime accidents and incidents involving suicide or criminal acts. Such detailed information has only become available very recently – that is, well after the dates of publication of the three MCA papers under consideration – and then only in relation to a very limited time period. This is another example of how the lack of relevant data limits the nature and extent of the performance analysis that one can do. (In any case, it is not yet clear to what extent the results from our analysis will have to be revised if and when such detailed information does become more generally available.)

2.5 Data collection problems in the OR study relating to the Musculo-Skeletal Unit

In the MSU-DES study, the main data involved the number of patients flowing through the musculo-skeletal service system, the types of specialist treatment that they required (and the activity times involved) and the availability of human and capital resources necessary to provide that treatment. Primary data were collected by members of the research team in several waves directly from the administrative staff of the Unit and secondary data from the administration systems of the hospitals involved. Additional information was gathered in interviews with senior members of the medical and nursing staff of the Unit, and by direct observation of its activities at various times during the study.

We soon became aware of frequent disputes between the senior staff of the Unit and managers further up the managerial hierarchy of the relevant NHS Trust about the reliability of such key data as the number of patients treated in a given period of time. Whereas we had tended to rely in the beginning mainly on secondary data from the hospital administration systems, we decided to collect the necessary information for our successive simulation models directly from the Unit's own administrative staff as the study progressed. (From direct observation, we knew that the Unit's members of staff were doing their best to record the patient throughput data as accurately as possible, while – for whatever reason – that did not necessarily appear to be the case for the hospital administration systems.)

Although issues concerning the reliability of patient data had not come as a complete surprise, we had not anticipated that data on the precise extent of the availability of medical and nursing staff might also give us problems. In the course of the study it became clear that staff availability could change at short notice, not just unintentionally through unavoidable staff absences (as was captured in our simulation models) but also, and more surprisingly, intentionally through extra resources being provided to cope with particularly heavy patient loads (of which we might only be told in general terms some time afterwards). We shall return to this last point in Sections 3.6 and 7.4.

3 ORGANISATIONAL PROBLEMS AND CHALLENGES

3.1 Organisational problems and challenges for performance measurement in public service organisations

When beginning our detailed analysis of the research situation and the particular structure of relations and constraints that maintain it, a review of the relevant academic literature throws up a number of specific problems and challenges for performance measurement in public service organisations. The main problems and challenges are as follows:

1. Multilevel organisational decision making
2. Multiple stakeholders
3. Multiple objectives
4. Multiple outputs and outcomes
5. The political nature of organisational decision making

We shall now briefly discuss each of these in turn. First, all public service organisations of the kind considered here are organised hierarchically (Jacobs *et al.*, 2006). Broadly speaking, organisational performance may be affected by decisions taken at its macro, meso and micro levels. Therefore, when analysing the performance of the decision making unit (DMU) under consideration, we should take account of the potential impact of decisions taken at higher or lower organisational levels.

Second, all public service organisations have multiple stakeholders (Dixit, 2002; Propper & Wilson, 2003; Bird *et al.*, 2005; Vissers & Beech, 2005). The three key stakeholders in any public service can be thought of as the funders or controllers of the service, the (professional) staff providing it, and the service users (or clients) (Ackroyd *et al.*, 1989; Brignall & Modell, 2000; Ackroyd *et al.*,2007). But the list of stakeholders is typically rather longer than that. For instance, Smith (2005) identifies six important groups of stakeholders in the NHS: patients (who are themselves a heterogeneous group), health care professionals, regulators (who come in many forms), taxpayers, the UK Government, and managers. The main problem is that these different stakeholders tend to have their own, differing views on the performance of the public service organisation concerned.

Third, all public service organisations have multiple objectives and there is typically a lack of consensus between the different stakeholders about the choice of objectives to be measured (in terms of outputs or outcomes) and the valuations or weights to be placed on these objectives (Van Peursem *et al.*, 1995; Smith, 1995b; Stone, 2002; Williams 2003; Smith & Street, 2005). Van Peursem (*op. cit.*, p. 61) expresses the problem for analysts and researchers (including OR practitioners) as follows:

> *"[T]he final choice of the measures deemed useful does not lie with us, but remains a social and political choice reflecting the value judgements of the community. Our concern is to enable that choice to be exercised from an informed position."*

More generally, Williams (*op. cit.*, p.654) notes that decisions about the trade-off between service productivity and service outcomes may cause intractable disputes at the political level:

> *"The tension between productivity and outcomes has not been resolved. ... Ignoring productivity is unrealistic in an environment of scarce resources. And ignoring outcomes would leave government programmes at risk of being pointless. ... Thus, in the political environment, champions of results ask for more resources, while champions of taxpayers assert that organisations need to make better use of the resources they have."*

Fourth, attempts by analysts and researchers to capture the multiplicity of outputs and outcomes in composite performance indicators are not necessarily helpful in this respect (Smith, 2002; Jacobs *et al.*, 2006; Jacobs & Goddard, 2007). Using Frontier

Analysis techniques, such as SFA and DEA, to construct composite indicators of efficiency (or cost-effectiveness) does not generally resolve the problem that we have just identified. Smith (2002, p. 423) argues:

> *"In my view the choices of outputs and weights [in a SFA or DEA model] are essentially political rather than technical problems. Political choices in this domain can of course be guided by technical analysis, in particular the use of well-designed population surveys. But in the end it is the job of elected politicians, rather than statisticians, to reconcile the often diverse popular views concerning the objectives and priorities of public services."*

Moreover, the use of composite performance indicators to construct league tables may mislead rather than enlighten public opinion. As noted by Jacobs & Goddard (*op. cit.*), the ranking of different DMU's in a league table is significantly influenced by the choice of decision rules on how to score the performance of different DMU's and the choice of weights applied to these scores – and these choices may well be ad hoc and arbitrary . Also, to disentangle genuine performance variations from random fluctuations, league tables ought to be published with indications of uncertainty to communicate the sensitivity of the reported measure – contrary to much of current practice.

Fifth, the overall nature of decision making with respect to performance measurement in public service organisations is highly political (Blundell & Murdock, 1997; Brignall & Modell, 2000; Williams, 2003; Cavalluzzo & Ittner, 2004; Smith & Street 2005). Blundell & Murdock (*op. cit.*, pp. 246-247) summarise the limitations to performance management in the public sector as follows.

> *"Despite the views of some people who would try to present performance management as a 'rational' managerial activity, it is still essentially a political process which is not without its dangers."*

Williams (*op. cit.*, p. 653) points to the political constraints on performance measurement in the public sector.

> *"Measurement arises in a political and social context, which both enables its occurrence and limits how it can develop. Proponents of performance measurement must build the political support necessary for long-term viability. To avoid misuse of performance reports, users must be aware of the political constraints that determine what and how information is reported."*

From this, Williams (*ibid.*, p.655) concludes that:

> *"The point is that both the practitioner and the end user of performance measurement should attend to the political constraints that lead to the particular form of performance reports. As with its analytic cousin, cost-benefit analysis, performance measurement may provide a false sense of objectivity when political constraints are ignored."*

In the UK, as in a number of other OECD countries, the rise of a set of doctrines for public sector management encapsulated in the concept of the 'New Public Management' (NPM) has led to dramatic shifts in the political constraints on performance measurement in public service organisations from the 1980's onwards (Hood, 1995 and 2007; Ferlie *et al.*, 2003; Bird *et al.*, 2005; Hodgson *et al.*, 2007; Ackroyd *et al.*, 2007). Hood (1995, p. 96) identified seven 'doctrinal components' of the NPM:

(i) Unbundling the public sector into corporatized units organised by product.
(ii) More contract-based competitive provision, with internal markets and term contracts.
(iii) Stress on private-sector styles of management practice.
(iv) More stress on discipline and frugality in resource use.
(v) More emphasis on visible hands-on top management.

(vi) Explicit formal measurable standards and measures of performance and success.

(vii) Greater emphasis on output controls.

In the 2000's, the NPM approach in the UK (and particularly England) seems to have culminated in 'a system of governance of public services that combined targets with an element of terror' (Bevan & Hood, 2006). But that kind of system is not problem-free, as stressed by Bevan & Hood (*ibid.*, p. 533):

> *"[A]lthough there were indeed dramatic improvements in reported performance, we do not know the extent to which these were genuine or offset by gaming that resulted in performance that was not captured by targets."*

We shall return to a fuller discussion of this latter problem in Section 5.1 below.

In the four papers under consideration, we are only dealing with two specific areas of public service. That is, we cannot claim to provide a comprehensive overview of the practical effects that the problems and challenges discussed in the present section have had on the UK public sector as a whole. However, in the next two sections of this chapter we shall explore the relevance of all of these problems and challenges, including any impact from the NPM, in the two areas of public service considered in our papers; namely, the coordination of Search And Rescue operations by the MSA and the provision of health care by the MSU.

3.2 Organisational problems and challenges for performance measurement in the Maritime and Coastguard Agency

In the MCA papers, the individual Coastguard coordination centres are treated as the main DMU's. However, the performance of these centres (in terms of the key outcome of the proportion of lives saved) is likely to be strongly influenced by decisions taken at MCA headquarters in Southampton. In particular, the overall management of the MCA decides on the appropriate staffing levels for each coordination centre and – more controversially – on which, if any, centres should be considered for closure. Of course, MCA management is itself subject to strong political pressures from the UK Government in this respect. At the micro level, on the other hand, the outcome of particular incidents depends, at least to some extent, on the quality of the decision making by individual watchkeeping officers on duty and the teams of local volunteers providing the SAR response.

Like other public service organisations, the Coastguard service has multiple stakeholders. In fact, the list of stakeholders is rather long and includes: the service users (who are themselves a heterogeneous group including seafarers, holiday makers, etc.), teams of local volunteers providing the SAR response, watchkeeping staff in the coordination centres, the PCS trade union representing many of the watchkeeping officers, MCA management in Southampton, the Department for Transport as the government department directly responsible for the service, and Members of Parliament (MP's) sitting on the House of Commons (HoC) Transport Committee, as well as other organisations and individuals having a stake in the service. The MCA-LRA paper highlights differences in opinion between some of these stakeholders, involving the Department for Transport and MCA management on one side of the argument and MP's on the HoC Transport Committee (and the PCS trade union) on the other side. Such disagreements have not been fully resolved to this day.

Although the Coastguard service has multiple objectives (rendering assistance in non-life-threatening incidents as well as life-threatening ones, educating members of the public etc.), there seems to be little disagreement between stakeholders that the key outcome is the proportion of lives saved. Similarly, none of the parties appears to be willing to defend any trade-off between that key outcome and (labour) productivity. Instead, the argument is couched in somewhat different terms.

In his 1999 report Lord Donaldson argued that, within limits, there is a positive correlation between labour productivity and the proportion of lives saved. This is based on the argument – familiar to other emergency services – that the watchkeeping officers in any particular coordination centre should have a sufficient workload to practise their skills regularly and not to become bored. However, if staff workload is increased by closing a coordination centre and dividing the workload between the adjacent ones, then the capacity of staff to hold the necessary level of 'local knowledge' may be impaired. Therefore, the closure of a coordination centre may entail a trade-off between the level of staff experience (which may be increased by the closure, with a positive effect on the proportion of lives saved) and their level of local knowledge (which may be reduced by the closure, with a negative effect on the proportion of lives saved). In general, the PCS trade union representing watchkeeping officers has argued strongly that the negative effects of centre closures have tended to outweigh any positive effects, with MCA management arguing the precise opposite.

There is no particular problem regarding the choices of outputs and weights in the SFA and DEA models that we used to construct composite performance indicators for the Coastguard service. In these models we tried to determine the effectiveness of different input combinations (the average scale of incidents, the average workload of watchkeeping staff dealing with these incidents, and the length of coastline monitored by the watchkeeping staff as a proxy variable for the amount of local knowledge required) for achieving a given level of the key outcome (the proportion of lives saved). As noted by Foreman-Peck (2002), that is a matter of what works in different circumstances, not a question of relative values.

The UK Government's shift towards the NPM approach has affected the political constraints on performance measurement in the Coastguard service. However, its impact was perhaps not as obvious as in other emergency services such as the ambulance service (Heath & Radcliffe, 2007) and the police service (Collier, 2006). In the ambulance service, the key performance indicators relate to measures of output – namely, response times – rather than outcome. According to some newspaper commentators – in particular, McIntosh (2005), as quoted in Heath & Radcliffe (*op. cit.*) – the Government's enthusiasm for the NPM approach, with its greater emphasis on output rather than process controls, has led some ambulance trusts to manipulate performance results in their favour (for instance, by changing the timing of the start or the completion of the response or by minimising the number of cases regarded as immediately life-threatening). Such accusations have received some support from an independent report into the performance of the ambulance service (CHI, 2003). In contrast, systematic manipulation of the key outcome results of the Coastguard service – namely, the proportion of lives saved of all those involved in life-threatening incidents – would be harder to accomplish, although not necessarily impossible. (This latter issue could only be resolved through a detailed investigation into the reporting of individual SAR incidents but – as already noted – we have had considerable problems in getting access to individual incident records.)

As far as the impact of the NPM approach on the police service is concerned, Collier (2006) notes that its orientation towards greater accountability has led to a shift since 2004 from process indicators in favour of outputs (detection of crimes) and outcomes (surveys of public satisfaction). In the present context, Collier's argument that effective performance measurement is hindered by insufficient understanding at the political level of cause-and-effect relationships between different means of policy intervention and outcomes would appear to be as relevant to the Coastguard service as to the police service. Quoting Sanderson (2001), Collier contends that the development of performance indicators has been primarily top-down with a dominant concern for enhancing control and upwards accountability rather than promoting learning and improvement. Again, this contention could also be applicable to the Coastguard service – at least according to some of the stakeholders concerned (such as the members of House of Commons Transport Committee (HoC Transport Committee SAR Report, 2005).

3.3 Organisational problems and challenges for performance measurement in the Musculo-Skeletal Unit

In the MSU-DES paper, the main DMU is the Musculo-Skeletal Unit itself. However, the performance of the Unit (in terms of the key outputs of the level of service that it provides to patients and the efficiency with which it utilises the available inputs) is likely to be strongly influenced by decisions at the level of the relevant NHS Trust (that is, the North Glasgow University Hospitals NHS Trust) and, ultimately, the Minister in charge of NHS Scotland. In particular, the management of the Trust determines the overall level of resources available – especially the number and type of medical and nursing staff working in the Unit. At the micro level, on the other hand, individual hospital consultants play a key role in managing the waiting lists of patients requiring their specialist services.

As we have seen, Smith (2005) identified six generic groups of stakeholders in the NHS. In the case of the Musculo-Skeletal Unit, its day-to-day activities were coordinated by senior members of the medical and nursing staff, although – as would be expected – Trust managers maintained tight control over the purse strings. We have already noted in Section 2.5 that there seemed to be frequent disputes between the senior staff of the Unit and relevant Trust managers about the reliability of performance data held at different organisational levels and probably about other important managerial issues as well. As usual in the NHS, the voice of the patients in the decision-making process was rather muted. (Although the Unit's medical and nursing staff tended to regard themselves as the main champions of their patients' interests.)

From an OR perspective, the Unit represents a queuing system, for which the two broad categories of outputs are, on the one hand, the number of patients treated over a given time period and the length of patient waiting times and, on the other hand, the efficiency with which the resource inputs are utilised. Standard queuing theory would lead one to expect a fundamental trade-off between these outputs: as the number of patients increases and the staff treating them become more & more busy, patient waiting times will tend to go up. While one would expect little disagreement between different groups of stakeholders about such basic features, that would not be true for the priorities allocated to the different outputs.

Patients would primarily be interested in shorter waiting times (and improved quality of health care – but that was not measured), while the staff of the Unit regarded increasing the total number of patients treated as an additional key objective. The senior staff of the Unit wanted to resolve the expected trade-off between patient numbers and waiting times by organising the Unit's activities (and utilising its resources) more efficiently – as described in some detail in the MSU-DES paper – but also by persuading Trust management to provide more resources in the form of additional staff. The management of the NHS Trust would be expected try and strike a balance between all of the outputs – number of patients of treated, waiting times and resource utilisation – but, given the inevitable funding constraints imposed at national level, there was some suspicion on the part of the Unit's staff that economising on resource usage was one of their main concerns in practice. The MSU study did not involve any attempt on the part of ourselves as OR practitioners to capture these various outputs in a composite performance indicator. Instead, the simulation model allowed each of the outputs to be analysed separately.

As noted by Smith (2005) and Bevan & Hood (2006), the UK Government's shift towards the NPM approach has had a big effect on the management of the NHS overall. However, because of its relative autonomy, the impact on NHS Scotland has been less pronounced than in other parts of the UK. Nevertheless, it would be surprising if NHS Scotland in general, and the organisational environment of the MSU in particular, had completely escaped the unintended and adverse consequences mentioned by Smith and Bevan & Hood, either in terms of distorted behaviour by NHS managers and staff and/or ineffective responses to patient needs in order to meet challenging performance targets or in terms of the manipulation of performance results for the same reason. Having said that, we found no evidence in our study that the MSU's own staff engaged in that kind of dysfunctional behaviour.

3.4 Some specific measurement problems for public service organisations

As part of our detailed analysis of the research situation, we should take account of some specific problems in measuring the performance of public service organisations. In Chapter 8 of their 2006 textbook, Jacobs *et al.* list four 'unresolved issues and challenges in efficiency measurement'; namely:

1. Output weights
2. Modelling the production process
3. Environmental constraints
4. Dynamic effects

In Sections 3.1-3.3, we have already discussed at some length the essentially political nature of decisions regarding the weights to be given to multiple outputs and outcomes. On the other hand, important questions about how to model the production process (i.e. the process transforming inputs into outputs and outcomes) will be dealt with in Chapter 5.

Environmental constraints and dynamic effects have also already been mentioned, albeit briefly. When measuring the performance of a public service organisation, one should try to take account of factors that lie outside the organisation's control. Figure 2 (in Section 2.1) provides a general overview of such 'environmental factors'. More specifically, they may include differences in (Jacobs *et al.*, 2006): relevant characteristics of the service users; features of the external environment such as geography, culture and economic conditions; activities of related agencies; the quality of the available resource inputs; and accounting treatments for valuing the inputs and outputs. If considered to be important, environmental factors may require suitable 'risk' or 'prior status' or 'case-mix' adjustments to be applied to the raw performance measures.

Figure 2 also highlights the importance of dynamic effects. These include the effects on current performance of the organisation's endowments, which are based on investments (or the lack of them) bequeathed by previous managers; but also the effects on future performance of investments made by the current managers of the organisation. Studying the organisation's performance over too short a period of time may cause one to miss such dynamic effects, thereby either overestimating current performance (if the current management has benefitted from previous investments in the service but does not make enough investments of its own) or, vice versa, to underestimate it. On the other hand, focusing on an excessively long time period means that the nature of the organisation, or the resources available to it, may have changed so much over time that one is no longer comparing like with like.

3.5 Measurement problems in the OR studies relating to the Maritime and Coastguard Agency

In the MCA-LRA paper, a key finding from our regression analysis was that the scale of incidents (measured as the average number of persons at risk for loss of life per 100 incidents) is an important environmental factor. That is, variations in the scale of incidents (which are not under the control of Coastguard management or watchkeeping staff) have a significant effect on the proportion of lives saved: the larger the incident, the higher the odds of success are likely to be. The inclusion of this factor as an explanatory variable in our regression model means that the performance indicator for each coordination centre is adjusted for its influence. But there may be other influences related to the varying nature of incidents that are not fully captured in this way. In particular, and as already noted in Section 2.4, until very recently we were unable to distinguish in our analysis between 'genuine' maritime accidents (which can be of any size) and incidents involving suicide or criminal acts (which are typically of a smaller scale). Only when such detailed data become more generally available can we ensure that the performance indicator for each coordination centre is fully adjusted for all such influences.

In the first couple of MCA papers (MCA-LRA and MCA-DEA), we based our analysis on annual SAR statistics for the relatively short period of 1995-98 (with 1999 statistics used for prediction tests on the regression model). However, in the next paper on this subject (MCA-BBN) we added statistics for 2000 – 2004 to our data set. While increasing the number of observations in this way is usually considered to be highly advantageous for statistical (estimation and testing) purposes, questions begin to arise about whether changes in the nature of the organisation or the quality of its available resources are adequately captured by the set of explanatory variables included in the model. We have reported some tests of this kind in the MCA-BBN paper, but further analysis will need to be done. (Again, this would be greatly facilitated by more detailed data becoming available.)

3.6 Measurement problems in the OR study relating to the Musculo-Skeletal Unit

In the MSU study, uncontrollable fluctuations in the patient mix would be expected to have a strong effect on all of the Unit's measurable outputs: the overall number of patients treated, the length of patient waiting times and the efficiency with which the available inputs are utilised. But as the simulation model was specifically designed to incorporate random variations in the patient mix, at least this environmental factor was fully accounted for in the model.

On the other hand, we became aware of dynamic effects that were less straightforward to deal with. These effects were not related to the length of the data set, because each of the individual projects discussed in the MSU-DES paper was relatively short-term and the simulation model was adjusted in the course of each successive project to take account of changing circumstances. But, as already noted in Section 2.5, it became clear that staff availability could change at short notice to deal with particularly heavy patient loads. In OR terms, this could be described as a negative feedback effect (equivalent to a control loop in a System Dynamics model), whereby an increase in a waiting list called forth a organisational response, in the form of a temporary injection of additional staff time, to reduce that waiting list again. If, as tended to happen, we were not accurately informed of this strictly temporary rise in resource availability, then the system appeared to cope with some types of demand increases, even though – according to the predictions of our simulation model – it 'should not' have been able to cope. As we shall discuss in Section 7.4, this caused obvious problems for validating our model. (Note that this issue only fully came to light after the MSU-DES paper had been written.)

4 THE PURPOSES OF THE OR MODELS

4.1 Managerial purposes behind performance measurement in public service organisations

We have already referred to the three broad aims of public service performance analysis in Section 2.1; these aims include providing a solid basis for public accountability by government officials, for instance. But as Pidd (2003) indicates, an OR model is usually developed with one (or more) specific client(s) in mind who, in the present context, may well be involved in managing the particular public service under scrutiny. Therefore, as well as paying attention to the broad aims of performance measurement and analysis, we need to consider the specific aims of any clients of the particular project that we are engaged in as OR practitioners.

According to Behn (2003), public service managers may want to achieve eight specific managerial purposes; namely to: evaluate; control; budget; motivate; promote; celebrate; learn; and improve. To control the performance of the public service is often seen as one of the most important managerial purposes (Boland & Fowler, 2000; Cavalluzzo & Ittner, 2004). But Behn contends that to foster improvement is actually the core purpose behind the other seven. This suggests a need to explore the role that OR models can play in helping managers (and other stakeholders) to achieve any desired improvements in the performance of a public service. Specifically, as we shall discuss in Chapter 5, this entails a more detailed investigation of the 'black box' of the process transforming inputs into outputs and, eventually, into outcomes (see Figure 1 in Section 2.1). But more generally, we should first analyse the different meanings of an OR model and the different motivations that have been suggested for developing OR models.

4.2 Different motivations for OR modelling

Pidd (*op. cit.*, p. 12) has proposed the following generic definition of an (OR) model:

> *"A model is an external and explicit representation of part of reality as seen by the people who wish to use that model to understand, to change, to manage and to control that part of reality."*

We may compare this to a generic definition of an (organisational) theory provided by Bacharach (1989, p. 496):

> *"A theory is a statement of relations among concepts within a set of boundary assumptions and constraints. It is no more than a linguistic device used to organise a complex empirical world. As Hall and Lindzey (1957, p. 9) pointed out, the function of a theory 'is that of preventing the observer from being dazzled by the full-blown complexity of natural or concrete events.' Therefore, the function of a theory is twofold: to organise (parsimoniously) and to communicate (clearly)."*

Despite the differences in wording, these two definitions have strong similarities. Pidd talks about a representation of reality, whereas Bacharach talks about a device to organise (our understanding of) reality. There would appear to be no great difference there, since one of the main reasons for representing part of reality by a model is, according to Pidd, to understand that part of reality better. Pidd's focus on part of reality corresponds to Bacharach's boundary assumptions and constraints. Similarly, Bacharach's argument that a theory should be parsimonious corresponds to Pidd's argument that a model is always a simplification of reality, albeit one intended for some definite purpose. (Bacharach's view of a theory as a linguistic device, seemingly excluding theories expressed in mathematical etc. terms, would appear to be unnecessarily narrow; and we may therefore ignore that particular contention.)

There is a potential difference between the two definitions in that Pidd mentions reality 'as seen by the people etc.' This may indicate a leaning by Pidd towards an interpretivist or constructivist epistemological stance (according to which the object of scientific study is not reality itself, but people's interpretations of it), which is not necessarily present in Bacharach's definition. (But we shall leave a full discussion of such issues to Chapter 7.)

Arguably a more important difference between the two definitions relates to the purpose of constructing the model or developing the theory. For Bacharach, the function of a theory is to help us in (mentally) organising reality and in communicating this organising framework to others. But Pidd regards a model as a potential basis for managerial intervention. More specifically, an OR model is not just a theoretical device for understanding the (social) world but is expected to lead to some sort of practical intervention into that world.

For some commentators, this latter distinction is crucial. For instance, according to Raitt (1979) as quoted in Ormerod (2006a, p. 905):

> *"The distinctive feature of OR in my view is its use of models and analogies. A model is not a theory, it has no direct substantive implications. No-one expects the history of OR to show an accumulation of models of increasing power, precision or generality. A model is constructed for practical application in a particular situation. It is wholly instrumental."*

Although we may regard Raitt's view as rather extreme, it does betray important differences between the philosophical positions that could be used to underpin the concept of a model. These different positions could be briefly summarised as follows:

1. *"Theory is indispensible. All research is necessarily theoretical even if only in the minimal sense that it is guided by prior conceptualisations of some sort."* (Fleetwood & Ackroyd, 2000, p. 12)

According to the first position, modelling should be placed in the service of theory development. Practical interventions should be informed and guided by relevant theories. On the contrary, the second position turns this argument on its head:

2. *"Pragmatism places theory in the service of practice. ... [I]t is difficult to see how the development of better algorithms or problem-solving methods has some higher purpose other than aiding practice in appropriate circumstances. ... However, pragmatism is not against theory, it simply gives it a different, less exalted role. According to pragmatism, theory should come out of a requirement of practice / action."* (Ormerod, 2006a, p. 906)

The first quote is potentially compatible with a range of philosophical positions; but this particular quote is taken from a textbook discussing a paradigm called 'Critical Realism'. The second quote is based on an alternative paradigm called 'Pragmatism'. The application of Critical Realism to OR modelling has been strongly championed by John Mingers (2000a; 2006a), but this is probably still a minority view among OR practitioners. Pragmatism, on the other hand, would currently appear to be one of the most popular paradigms in OR (Ormerod, 2006a). (Although it will be argued in Chapter 7 that Critical Realism has certain advantages over Pragmatism, we shall leave that particular debate for the moment.)

The difference between the two positions outlined above may be taken to correspond, at least roughly, to Mitchell's (1993, p. 113) distinction between:

1. The idea of a model as a statement of the beliefs held, including assumptions willingly made, relevant to the issue under study.
2. The idea of a model as a device (usually for making predictions).

According to Mitchell, these two different meanings correspond to two key motivations of much OR work, respectively:

1. Belief creation and codification to secure understanding and to allow manipulation of better beliefs.
2. Belief manipulation and its pragmatic application to helping select suitable problem-solving actions.

Note that Mitchell's distinction is not dissimilar to Roy's (1993) distinction between OR as a 'decision science' and OR as a 'decision-aid science'. We might say that the first type of model represents OR as a science to aid generalisable understanding (or theory development) and the second type of model represents OR as a science to aid problem solving. Each of these two different types of model may play a role, albeit perhaps a different one, in helping public-service managers (as the clients of a relevant OR project) to achieve their core purpose of fostering improvement – cf. the INFORMS slogan for OR as the 'science of better' (www.scienceofbetter.org) – and also, more generally, to support the three broad aims of public service performance analysis.

4.3 The purpose of the OR studies relating to the Maritime and Coastguard Agency

In the MCA papers, the OR models that we developed are similar in function to Bacharach's (1989)'s concept of a theory; namely, as a way of mentally organising reality and in communicating this organising framework to others. Or, looking at the issue more from an OR perspective, the models correspond to Mitchell's first type of model: a statement of the beliefs held for securing better understanding. In other words, the models fit in with the idea of OR as a science to aid theory development.

When we started to develop models to measure and analyse the performance of Coastguard coordination centres, we did not have any specific clients in mind. Our initial interest in this topic was raised by newspaper reports at the time (around 1999) about the decision by senior management of the MCA to close a number of coordination centres. As there seemed to be a considerable level of controversy surrounding this decision (as reviewed briefly in the MCA-LRA paper), we became interested in analysing the rationale for these closures and in gathering empirical evidence to support such analysis. While not taking an a priori position on whether the centre closures were justified or not, we wanted to expose such important managerial decisions about the organisation of the Coastguard service to public accountability – and to do so more effectively than the report by Lord Donaldson on the matter, which did not seem to be based on formal (scientific) modelling of an OR kind or otherwise.

In the course of our studies, we contacted various important stakeholders for information – first mainly about the closure decisions but somewhat later, as we widened our interest in the issue, also about the more general question of how best to measure and analyse the performance of the Coastguard service. In particular, we conducted interviews with watchkeeping staff in the coordination centres, officials of the PCS trade union (which represents many of the watchkeeping officers), and subsequently also the MP chairing the House of Commons (HoC) Transport Committee.

The MCA management in Southampton was the only party to show any reluctance in actively cooperating with us in our research. We have never been given any reason for this (at least not by MCA managers themselves). However, we received the impression that the MCA management was at first suspicious that the 'real' motive for our research was not so much to promote public accountability or anything like that, but rather to identify (and subsequently publicise) the functional competence (or lack of it) of the senior managers of the organisation.

We managed to gain more active involvement from MCA management in our research only in the last few years – after the centre closures had been put into effect and the political furore had calmed down a bit (although it has still not disappeared altogether). In September 2005 we were invited to give a presentation about our research at MCA headquarters in Southampton, in which we focused on how our research into the performance of the service might be used to achieve the other broad aim of performance measurement (that is, apart from public accountability and assessing functional competence); namely, to establish 'what works' and, in the course of doing so, to identify opportunities for performance improvement. From then on, the role of MCA management has moved some way towards becoming one of the clients for the overall project. (Other potential clients include the PCS trade union and those MP's taking an active interest in this whole issue.) But we cannot yet claim that our models appear to aid managerial decision making in any direct sense.

4.4 The purpose of the OR study relating to the Musculo-Skeletal Unit

In the MSU study, on the other hand, finding out what works well and where performance improvements can be made was both our own broad aim as OR practitioners and the specific managerial purpose of the senior medical and nursing staff who were our clients. That is, the simulation model that we developed to analyse patient numbers and waiting times corresponds more to Mitchell's second type of model: a device for helping to select suitable problem-solving actions. In other words, the MSU simulation model fits in well with the idea of OR as a science to aid problem solving.

From the start, the OR study was intended to help the senior members of the medical and nursing staff to gain a better understanding of the main performance drivers of the Musculo-Skeletal Unit and to predict (on the basis of simulated experiments) the effects of potential actions to improve its performance. As explained in the MSU-DES paper, the form and content of the successive versions of the simulation model was very much determined by the specific questions that our clients wanted to ask at each stage of the study. Therefore, unlike the MCA case, the modelling process in the MSU study has been characterised by close interaction with, and participation from, our clients. In short, our motive was to facilitate organisational learning and thereby to aid managerial decision making.

5 MODELLING THE TRANSFORMATION PROCESS

5.1 Modelling the transformation process and understanding cause-and-effect relationships in public service organisations

Appreciation of the research situation and analysis of the information flowing from it should lead OR practitioners to assessing potential explanations and, if appropriate, considering ways in which organisational performance could be improved. As we have discussed in Chapter 4, Behn (2003) argues that fostering improvement is the core purpose of public service managers. More generally, we have already repeatedly referred to the need to establish what works in promoting stated objectives as one of three broad aims of public service performance analysis (Bird *et al.*, 2005). Therefore, we should not just focus on the accurate measurement of performance in terms of the four E's, but also try to open up the black box of the transformation process shown in Figure 1. Using Mitchell's (1993) terms, we should develop 'structural' models (based on the identification of cause-and-effect relationships) rather than – or as well as – black-box models.

However, Behn (*op. cit.*, pp. 597-598) notes how complex the transformation process is and, therefore, how difficult it is to predict the results of managerial interventions.

> *"To ratchet up performance, public managers need to understand how they can influence the behaviour of the people inside their agency (and its collaboratives) who produce their outputs and how they can influence the conduct of citizens who convert these outputs into outcomes. ... They need to know what is going on inside their entire, operational black box. ... Unfortunately, what is really going on inside the black box of any public agency is both complex and difficult to perceive. ... It is very difficult to understand the black box adjustments and interactions that happen when just a few of the inputs (or processes) are changed, let alone when many of them are changing simultaneously and perhaps in undetected ways."*

Although Bird *et al.* (*op. cit.*), referring to Propper & Wilson (2003), advocate the use of experiments to resolve a range of research questions relating to performance measurement in public service organisations, it would practically impossible for researchers to conduct 'controlled experiments' (i.e. experiments in an artificially closed system) in which a particular mechanism is observed in isolation. (Danermark *et al.* (2002) argue, following Harré (1970) and Bhaskar (1978), that the infeasibility of controlled experiments is a salient characteristic of the social sciences in general.) Conducting simulated experiments on the basis of a previously constructed simulation model – whether using discrete-event simulation, systems dynamics or some other form of simulation modelling – is an attempt to get round this problem; but its effectiveness depends crucially on the validity of the underlying simulation model. (We shall discuss model validation in depth in Chapter 7.)

One particular reason why 'ceteris paribus' conditions cannot be enforced in practice is – as many researchers have pointed out (Likierman, 1993; Smith, 1995a; Ferlie *et al.*, 2003; Propper & Wilson, 2003; Bird *et al.* 2005; Collier 2006; Bevan & Hood, 2006; Hood 2007; Heath & Radcliffe, 2007) – that managerial interventions may create 'perverse incentives' among the (professional) staff providing the service and thereby lead to unintended consequences (at least from the managerial or regulatory perspective). One of the chief underlying causes is that inefficiency is inherently unobservable and that estimates of efficiency have to be derived indirectly after taking account of observable phenomena (Jacobs *et al.*, 2006). But by prioritising those performance indicators that can be observed more easily, other organisational activities are marginalised (Van Peursem *et al.*, 1995) and, therefore, become susceptible to 'gaming' by public sector employees (Bevan & Hood, *op. cit.*). Bevan & Hood (*ibid.*, p. 521) define gaming as "reactive subversion such as 'hitting the [performance] target and missing the point' or reducing performance where targets do not apply." (But note that, as we have already explained in Sections 3.2 and 3.3 above, we did not find evidence in our own studies that staff from either the MCA or the MSU engaged in that kind of dysfunctional behaviour. Therefore, the issues of perverse incentives and/or gaming did not play a significant part in our OR models for these particular public service organisations.)

More generally, Boland & Fowler (2000) argue that performance management in public sector organisations usually takes the form of a causal loop that is established between perceived performance and resulting actions, thereby constituting a form of feedback control. As noted by Bevan & Hood (*op. cit.*), this is a form of homeostatic control requiring a negative feedback loop to keep performance on target (underperformance leading to managerial actions intended to improve performance). However, Boland & Fowler warn of the ever-present danger of initiating positive feedback loops with potentially disastrous consequences for overall behaviour or performance (underperformance unintentionally leading to actions from service providers that reduce performance even further). The problem is neatly summed up by Smith (1995b, p.16), referring to Osborne & Gaebler, 1993):

> *"The crucial question is what model of service delivery will minimise the incidence of dysfunctional behaviour, and persuade public sector employees to deliver excellent services."*

In line with Behn's (2003) argument, Hodgson *et al.* (2007) argued that the question of what works in bringing about public sector performance improvement does not have a straightforward answer. From a metastudy of 51 empirical studies (quantitative, qualitative and mixed-method) of improvement in public organisations, they found the quality of evidence for specific improvement 'triggers' to be limited and identified the following list of problems: diversity in research methods employed and service contexts considered; difficulty of isolating one or more effective triggers for improvement; dangers of taking 'snapshots' of improvement; lack of discussion of improvements in one area having a negative impact in another (i.e. trade-offs between performance objectives); research focus on improvements in processes rather than in outputs or outcomes.

Accordingly, Hodgson *et al.* (*op. cit.*) advocated sustained research in order to establish cause-and-effect relationships between the adoption of a particular approach to performance improvement and the likely results for different stakeholders. This apparent need for better theoretical explanations of what drives performance in public service organisations is widely recognised in the relevant academic literature. According to Ferlie *et al.* (2003, pp. S10-S11):

> *"Our analysis of the papers published in the British Journal of Management over the last decade shows that they share the characteristics of much UK social science in being more empirical than theoretical. ... A stronger focus on theory (while retaining a strong grounding in empirical evidence-gathering) will enable researchers not only to examine outputs and outcomes but also to develop an understanding of change processes which underlie the phenomena being examined. While it has been common practice to draw on the current UK Government's mantra of 'what matters is what works', for researchers this is not sufficient. We need to know what works, for whom, when, in what circumstances and why. ... Without a theoretical framework, however rudimentary, public management research will be vulnerable to the pressures and politics of the day and will be unable to create meaningful generalisable findings which are of wider relevance, across services, sectors and countries."*

Arguments for developing a sound theoretical basis for research in this area can also be found in various forms in papers by Brignall & Modell (2000), Thorpe & Beasley (2004), Smith (2005), Johnsen (2005), and Micheli *et al.* (2005) and, in the more general area of best-practice benchmarking, by Francis & Holloway (2007). Developing such a theoretical basis would involve conducting more longitudinal analysis (Brignall & Modell , *op. cit.*; Ferlie *et al.*, *op. cit.*), but also applying a wider set of academic disciplines (Ferlie *et al.*, *op. cit.*; Johnsen, *op. cit.*; Micheli *et al.*, *op. cit.*), matching methodological choices to the purpose of the analysis (Smith & Street, 2005) and triangulating data across different research methods (Ferlie *et al.*, *op. cit.*; Bird *et al.*, 2005; Micheli *et al.*, *op. cit.*). However, we shall return to such methodological issues in Chapters 6 and 7.

5.2 Different types of OR models

As first indicated in Chapter 4, and explained in more detail in Section 5.1 above, we want to examine the cause-and-effect relationships inside the black box of the process transforming inputs into outputs and outcomes in public service organisations. But before discussing which kind of OR models could be particularly useful in this respect, it will be helpful first to map the terrain more generally by reviewing briefly how OR models might differ in their salient characteristics. In his 1993 textbook on the practice of OR, George Mitchell devotes a chapter on models (namely Chapter 8). We have already quoted part of the beginning of that chapter in Section 4.2 above (on the different motivations for OR modelling). Further on in that chapter, Mitchell distinguishes between different types OR models according to seven relevant dimensions:

1. The level of abstraction and simplification involved.
2. A number of related dimensions, including:
    a. Structural versus 'black box'.
    b. Exploratory versus predictive.
    c. Realistic versus instrumental.
    d. Micro versus macro.
    e. The extent of the computing power required.
3. Standard versus purpose-built.
4. Absolute (valid anywhere and anytime) versus relative (situation-dependent).
5. A number of related dimensions, including:
    a. Passive (incorporating no beliefs about people's behaviour) versus normative (incorporating beliefs about the best action to be chosen).
    b. Passive (defined as before) versus behavioural (incorporating beliefs about how people will react).
    c. The extent to which a model is interactive (where the relevant beliefs are split into two; namely, beliefs for which a consensus exists among the group of decision makers and beliefs that the group prefers to review and reflect on before using it).

6. Private versus public.
7. Part (modelling a part of a larger system) versus whole (modelling the system as a whole).

We shall refer to Mitchell's distinction between different types of OR models in Sections 5.5 and 5.6 below, in which we discuss some basic characteristics of the OR models in our studies. Also, the difference between macro- and micro-level models will be an important facet of the discussion in Section 6.2.

5.3 Using OR models to understand cause-and-effect relationships in public service organisations

General lists of OR modelling approaches, as well details of particular methods, can be found in more or less all basic textbooks on Operational Research (or Management Science) – see Waters (1998) for a fairly representative example. In his 2003 textbook on modelling in Management Science (which we have already referred to in Sections 4.1 and 4.2), Pidd places the various modelling approaches in an empirically-based framework, based on the usual distinction between 'hard' and 'soft' OR methods. The hard OR methods are explained under the heading of "mathematical and logical modelling" and include Optimisation / Mathematical Programming, Heuristic Search Methods and Discrete Event Simulation; whereas, the soft OR methods are explained under the heading of "interpretive modelling" and include Soft Systems Methodology, Cognitive Mapping / SODA and System Dynamics.
(Apart from comparing different OR modelling approaches, Pidd also discusses more general principles of OR modelling, which we shall consider briefly in Section 5.4 below.)

The list of OR modelling approaches used in support of performance measurement is typically rather more restricted. For example, in their 2005 textbook on efficiency and productivity analysis, Coelli *et al.* focus on a limited range of methods; namely, index numbers and Data Envelopment Analysis (DEA), followed by econometric estimation (through regression analysis), in general, and Stochastic Frontier Analysis (SFA), in particular. Similarly, DEA and SFA constitute the main approaches discussed by Jacobs *et al.* in their 2006 textbook on efficiency measurement in public service organisations providing health care. These examples confirm the central role played by DEA and SFA models in public sector performance measurement, as we have already indicated towards the end of Section 2.1 above.

But given our desire to peer inside the black box of the transformation process in public service organisations, we need to complement Frontier Analysis techniques –

such as DEA and SFA – that may help us to uncover significant differences in performance, with other OR modelling approaches that enable us to gain a better understanding of the cause-and-effect relationships driving these performance differentials. Quoting the title of a recent textbook on OR modelling of health care systems (Sanderson & Gruen, 2006), we may call the latter approaches 'analytical models for decision making'. (Davenport seems to have coined the phrase 'decision analytics' in a similar context (Johnson, 2007).) In keeping with our discussion in Section 4.2 about the key motivations of OR modelling, the use of such analytical methods can support both client-focused problem solving based on existing theories and the development of new theoretical approaches.

In their review of OR methods to support management decision making in health care, Sanderson & Gruen (*op. cit.*, chapter 12) divide OR modelling approaches into three broad areas: (1) models for clarifying complex decisions; (2) models for planning and allocating resources; and (3) models for evaluating effects of changes in systems. This classification can arguably be applied more generally to OR modelling for public service organisations (and not just those engaged in the field of health care). Out of the three broad areas listed above, the third one – i.e. the "models for evaluating effects of changes in systems" – would then seem to be particularly useful for gaining a better understanding of what drives performance differentials relating to the same organisational unit at different points in time (as in our MSU study) or relating to different organisational units at the same or different points in time (as in our MCA studies). The specific kinds of OR methods listed by Sanderson & Gruen under this third heading include Markov models, System Dynamics, Queueing Theory and Discrete Event Simulation.

Rather than trying to apply all of the different methods under Sanderson & Gruen's third heading, we selected Discrete Event Simulation (DES) for our MSU study. We chose DES in favour of an alternative simulation approach such as System Dynamics, for instance, mainly because DES has long been considered as a standard approach in the relevant OR literature (e.g. Davies & Davies, 1994) for modelling the potentially complex interactions between patient flows and available resources and, even more importantly, this approach readily lends itself to conducting detailed experiments relating to potential interventions in the system – as explained in more detail in the MSU-DES paper and also Section 5.6 below.

In the context of our MCA studies, on the other hand, we have applied an alternative kind of analytical model – not (yet) covered in basic textbooks on OR modelling such as Sanderson & Gruen – in the form of a Bayesian Belief Net (BBN). While the reasons for selecting the BBN approach are explained in some detail in Section 6.2 below, it is worth noting at this point that the detailed empirical data on which to base micro-level simulation models (whether DES or System Dynamics), even if fairly rudimentary, were not available to us – leaving the construction of a BBN as one of our few practical modelling options at the micro level. (However, it is not impossible that a BBN may form a useful stepping stone in the development of a future micro-level simulation model of some kind.)

To sum up on the basis of our arguments so far, we may distinguish between four different categories of OR modelling approaches:

1. Models that are based on the application of problem structuring methods such as Cognitive Mapping (SODA or Journey Making) or Soft Systems Methodology. These would all be included in Pidd's list of soft OR methods; and they would fall under Sanderson & Gruen's first heading of "models for clarifying complex decisions".
2. Models (usually, but not always, deterministic) that involve the application of some kind of optimisation method, typically of a mathematical nature (e.g. mathematical programming or heuristic search methods). These would all be included in Pidd's list of hard OR methods; and they would fall under Sanderson & Gruen's second heading of "models for planning and allocating resources".
3. Models (usually, but not always, stochastic) that involve the application of some kind of simulation, or other non-optimising analytical, method. Some of these – such as Markov models, Queueing Theory or Discrete Event

Simulation – would normally be included in Pidd's list of hard OR methods; and others – such as System Dynamics and Bayesian Belief Nets – in his list of soft OR methods. All of these models, including arguably BBN's, would appear to fall under Sanderson & Gruen's third heading of "models for evaluating effects of changes in systems".

4. Models for measuring performance; or, more particularly, models for measuring performance differentials between decision-making units that are broadly comparable otherwise. Data Envelopment Analysis and Stochastic Frontier Analysis are important examples of these kinds of models. Their specific role would not seem to be adequately captured by Pidd's hard-soft dichotomy (although DEA may be seen as 'softer' than SFA as it seems to involve a greater degree of interpretive modelling). Also, these kinds of models are not considered by Sanderson & Gruen, although – if one had to include them in any of their three categories – the third heading of "models for evaluating effects of changes in systems" would appear to fit them best.

In the four papers considered in this thesis, we have focused on OR models in the third and fourth categories listed above. Rather than explicitly applying the methods included in the first category above, problem structuring issues were dealt with in the process of developing the initial regression models for the MCA studies (as discussed in the MCA-LRA paper and Sections 5.5 and 6.2 below) and the initial simulation models for the MSU study (as discussed in the MSU-DES paper and Section 5.6 below). The application of mathematical optimisation methods as included in the second category above, on the other hand, did not seem to be directly relevant to the particular issues that we were faced with in the MCA studies. With regard to the MSU study, any attempt at optimisation was on the basis of trial-and-error (based on the results of simulated experiments) rather than the application of formal mathematical methods. (That is not to say that the use of mathematical programming or heuristic search methods might not become appropriate at some future stage in either of these contexts.)

In the final two sections of this chapter we shall return to the specific OR models used in our studies. But first we shall briefly review some more general principles of OR modelling.

5.4 General principles of OR modelling

In Section 1.2 we listed five research questions for this thesis. The first of these questions – relating to the particular characteristics of performance measurement in public service organisations – has already been investigated at some length in Chapter 3 above. The last three questions are concerned with multimethodology as a modelling approach, which is the main focus of Chapters 6 and 7 below. In partial response to the second question – relating to the application of OR models in support of performance measurement in public service organisations – we aim to demonstrate in the present chapter how the choice of particular OR modelling approaches can be linked directly to the respective purposes of the MCA and MSU studies (as discussed in Chapter 4).

But the second research question is also concerned with the extent to which general principles of OR modelling are applicable within the particular context of public service performance measurement. In simple terms, the issue can be expressed as follows: even if we manage to select OR modelling approaches that are appropriate for the purpose in hand, are there more general guidelines that we should observe when applying our models – or are such general modelling principles of little value in practice?

A prominent candidate for a set of general principles of OR modelling is the list provided by
Pidd (2003), as summarised below:

1. Model simple, think complicated.
2. Be parsimonious, start small and add.
3. Divide and conquer, avoid mega models.
4. Use metaphors, analogies and similarities.
5. Do not fall in love with data (including: the model should drive the data, not vice versa).
6. Model building may feel like muddling through.

To respond to the remaining part of the second research question, we shall examine to what extent we have been able to apply Pidd's general principles in the development of the models for the four papers under consideration. But rather than attempting to do so in the present chapter, we shall return to this issue when we are in a position to take an overall, summary view of our modelling efforts; that is, in Chapter 9.

5.5 Understanding cause-and-effect relationships in the Maritime and Coastguard Agency studies

As we have already explained in Chapter 4, our initial interest in the performance of Coastguard coordination centres had been raised by newspaper reports about the proposed closures of some centres. As the purported aim of these closures was to improve the performance of the Coastguard service, we wished to gain a better understanding of which particular factors influence this performance (particularly in terms of the proportion of lives saved) and how these factors might be affected by the closures. To this purpose, we needed to investigate what causes performance variations between individual Coastguard coordination centres and to what extent the Coastguard service was becoming more effective, or less, in saving lives over time.

However, the nature of the process in which a heterogeneous set of inputs (various grades of watchkeeping officers and various types of capital resources) is transformed, in the presence of many potentially relevant environmental factors, into a range of outputs and ultimately, with the involvement of both local SAR teams and service users, into the specific outcome of saving lives could reasonably be expected to be rather complex. Moreover, our detailed literature review showed that this process had not been modelled before – at least not in any way that would help us to understand the potential impact of centre closures. The report by Lord Donaldson (1999) had identified a key trade-off between the level of staff experience (which may be increased by the closure of a centre as the workload is divided between the remaining ones, with a positive effect on the proportion of lives saved) and their level of local knowledge (which may be reduced by the closure, with a negative effect on the proportion of lives saved). But as Lord Donaldson's analysis had not been based on any kind of formal modelling approach, this did not provide us with the detailed understanding that we were looking for.

MCA management, (as opposed to the PCS trade union representing watchkeeping officers) tended to deny suggestions that centre closures would any have any significant impact on the amount of local knowledge possessed by watchkeeping officers in the remaining centres. Therefore, any loss of local knowledge that might occur as a result of managerial decisions to close certain centres would have to be regarded as an unintended consequence of these decisions. In consequence, one of our main aims was to model the effect that the amount of local knowledge required – for which we used the length of coastline monitored by watchkeeping staff as a proxy variable – tended to have on the performance of each centre, and how important this effect was in relation to other explanatory variables, both under the control of management (such as average staff workload) and not under its control (such as the average scale of incidents).

With respect to the OR models concerned, we primarily employed models in the third and fourth categories from our list of OR modelling approaches in Section 5.3 above. Applying Mitchell's (1993) terminology (see Section 5.2 above), these models were intended to be structural (rather than black box), exploratory (rather than predictive) and realistic (rather than instrumental). We started with multiple regression analysis to develop a 'macro' (aggregate) model for the transformation process underlying the annual Coastguard statistics. We were thus able to identify statistically the main explanatory factors influencing the performance of the service. Because our key outcome indicator is binary in nature – the life of a person at risk is either saved or lost – this first set of regression models was based on (Binary) Logistic Regression Analysis (LRA). Using our initial findings we then performed Frontier Analysis – both of the parametric (SFA) and the non-parametric (DEA) kind – to measure the 'efficiency' of individual coordination centres over time. That is, we estimated how close each centre was to the relevant optimal performance frontier. (Note that the establishment of a performance frontier involves some notion of optimisation.) These forms of analysis formed the main subjects of the MCA-LRA and MCA-DEA papers. Subsequently, we turned to the construction of Bayesian Belief Nets to gain a more detailed understanding of the cause-and-effect relationships affecting the outcomes of individual incidents. This 'micro' model formed the main subject of the MCA-BBN paper. We shall return to the interconnections between the different OR models in Section 6.2.

5.6 Understanding cause-and-effect relationships in the study of the Musculo-Skeletal Unit

The process by which the Musculo-Skeletal Unit transforms a wide range of human and capital resources into key outputs, as measured by the level of service provided to patients and the efficiency with which the inputs are utilised, is rather complex. Unlike the case of the MCA studies, the general features of this kind of transformation process were already reasonably well understood on the basis on similar studies done by OR practitioners elsewhere. (The MSU-DES paper contains only a small selection of the references that are potentially available.) But in order to explain where and why bottlenecks occurred in this particular Unit, we needed to investigate the context-specific factors that appeared to influence the number of patients seen and the length of waiting times experienced at various stages in the patients' journey through the Unit. As usual with discrete-event simulation model, the stochastic nature of the models related mainly to the number and nature of patient arrivals, the times needed to complete the various activities in the Unit, and the different routes that patients could take through these activities (depending on the nature of their medical condition and other relevant factors). Unlike some of the models that we developed in the MCA case, we made no attempt to establish an 'efficiency frontier' to compare explicitly the performance of this particular Unit to that of similar Units in other NHS hospitals.

However, opening up the black box of the transformation process and looking inside was only the first step. As the senior members of the medical and nursing staff of the Unit, our clients wanted to use the OR model to predict, with a reasonable degree of accuracy, the impact of potential interventions in the system. In particular, they wanted to know what actions would work – at that particular time and in their particular circumstances – to resolve the expected trade-off between patient numbers and waiting times. Each of the successive simulation models was therefore developed for the express purpose of conducting simulated experiments (as the nearest possible substitute for controlled experiments on the real system) in order to predict the range of consequences – whether intended or not – of more or less complex initiatives such as the implementation of a new patient triage system. As explained more fully in the MSU-DES paper, the overall modelling approach employed for this purpose was participatory, iterative and focused on enhancing the clients' understanding of the main performance drivers of the service.

As regards the OR modelling involved, we developed a sequence of non-optimising, stochastic (discrete-event) simulation models (that is, models in the third category from our list of OR modelling approaches in Section 5.3 above). In other words, we saw no need to adopt a multimethodology approach as in the MCA papers. Applying Mitchell's (1993) terminology (again, see Section 5.2 above), these simulation models were intended to be structural and realistic. But the models were more situation-dependent than in the case of the MCA studies and were meant to be able to predict (on the basis of simulated experiments) as accurately as possible the effects of potential actions to improve the Unit's performance. That is, the models were meant, first of all, to explain at a micro level the performance of the service system as configured at various points in time and, secondly, to enable experimentation with different system configurations.

6 MULTIMETHODOLOGY IN OR MODELLING

6.1 The role of multimethodology in OR modelling

In Section 5.1 we have referred to the numerous arguments in the recent academic literature about the need to develop a sound theoretical basis for research on performance measurement in public service organisations (for instance, Ferlie *et al.*, 2003). More particularly, we alluded to recommendations to apply a wider set of academic disciplines, to match methodological choices to the purpose of the analysis, and to triangulate data across different (qualitative and quantitative) research methods (Ferlie *et al.*, *ibid.*; Johnsen, 2005; Micheli *et al.*, 2005; Smith & Street, 2005; Bird *et al.*, 2005). Such recommendations towards taking a multimethodology approach in this area of research appear to work both ways. Social scientists who are familiar with the use of conventional (mainly qualitative) case study methods, such as Ferlie *et al.*, argue that these should be complemented by a wider set of methodological choices. And proponents of the application of positivist methods common in classical statistics (based on the formulation of hypotheses that can be statistically tested through appropriate experiments), such as Bird *et al.*, stress the need to engage also in qualitative studies.

As Mingers has argued forcefully in a series of papers over the last decade, multimethodology already plays an important role in more general OR modelling (Mingers & Brocklesby, 1997; Munro & Mingers, 2002). Before considering this issue in some more detail, it is helpful to review the definition of some key concepts. Mingers & Brocklesby (*op. cit.*, pp. 490-491) propose the following definition of a 'methodology'.

> *"A methodology is a structured set of guidelines or activities to assist people in undertaking research or intervention. Generally, a methodology will develop, either explicitly or implicitly, within a particular paradigm and will embody the philosophical assumptions and principles of the paradigm. Usually there is more than one methodology within a paradigm."*

Simply speaking, a (research) methodology is a set of (research) methods, where a 'method' may be defined as follows by (Mingers, 2006a, p. 214).

> *"A method is a well-defined sequence of activities that, if carried out proficiently, yield predictable results."*

Thus, according to Mingers & Brocklesby, the methodology specifies what type of activities should be undertaken, and methods are particular ways of performing these activities. A key point is, however, that a research methodology is typically associated with a particular scientific paradigm. Unfortunately, the term 'paradigm' does not have a single, universally-accepted, meaning. In writings on research philosophy, it first came to prominence through the well-known work of Kuhn (1962), who defined a paradigm as the 'shared beliefs in a research field'. Subsequently, Burrell & Morgan (1979) defined paradigms as 'tightly-coupled ideologies, ontologies, epistemologies and methodologies that guide modes of organisational analysis' – which is the kind of meaning that tends to be most commonly used in social science research nowadays. In OR modelling, an obvious example of a methodology would be Soft Systems Methodology (SSM), which – according to Checkland (2006) – embodies the assumptions and principles of phenomenology as its overarching paradigm.

We have already noted that a paradigm may encompass a set of different methodologies. For instance, phenomenology is taken by Checkland (*ibid.*) to encompass not just SSM, but also other 'soft' OR methodologies such as Strategic Choice (SC) and Strategic Options Development and Analysis (SODA). It is, therefore, important to distinguish between two different forms of multimethodology. The first – more straightforward – form involves mixing different methods within a single paradigm. But in the second – potentially more complex – form, methods are mixed across different paradigms. (We shall explain shortly why such 'multiparadigm' research is often considered to be especially problematic.)

Mingers & Brocklesby (*op. cit.*) and Mingers (2003) have suggested a number of arguments in favour of a multimethodology approach to underpin OR modelling:

1. Real-world problem situations are highly complex and multidimensional. Different methodologies focus on different aspects of the situation.
2. Different methodologies may be used in different phases of modelling process.
3. Different methodologies are already widely combined in practice – usually within a kind of paradigm that is based on Pragmatism (cf. Tashakkori & Teddlie, 1998; Ormerod, 2006a) or a Postmodernist perspective (cf. Schultz & Hatch, 1996; Spender, 1998).
4. Combining different methods can improve the reliability of research results through triangulation.

(Note that these arguments broadly mirror the recommendations towards taking a multimethodology approach in research on performance measurement in public service organisations that we discussed earlier.) In order to investigate the extent to which multimethodology is actually practised, Munro & Mingers (2002) conducted a survey of (both academic and non-academic) OR practitioners. From this, they drew the following conclusions: multimethodology is common in OR, 'perhaps increasingly so'; it is judged to be very successful by those who practise it; most combinations are either of 'hard' or 'soft' OR methods, rather than of hard and soft together; methodological issues do not seem to get much attention from the OR practitioners involved; and many OR practitioners now consider themselves 'multidisciplinary' (which would fit in with Tranfield & Starkey's (1998) concept of the 'mode 2' method of knowledge production).

As for academic papers, Mingers & Brocklesby (1997) refer to a number of articles published up to the mid-1990's that involve some kind of multimethodology in OR modelling. More recently, many further examples have appeared, including: the proposal by Mingers (2000b) of a practical approach to the process of combing 'soft' and 'hard' OR methods; a range of practical examples of multimethodology interventions given in the well-known book edited by Rosenhead & Mingers (2001), particularly in the chapter by Ormerod; a series of papers by Eden and various research collaborators (e.g. Howick *et al.*, 2008) on successfully combining different OR methodologies such as Cognitive Mapping and System Dynamics; and a recent paper by Kotiadis & Mingers (2006) describing a combination of a soft OR methodology (SSM) with another methodology (DES) that is commonly seen as towards the harder end of the OR spectrum.

From the above discussion, we may safely conclude that there is a distinct role for multimethodology in OR modelling as applied specifically to performance measurement in public service organisations. However, this should be done in order to satisfy the need for a sound theoretical basis in this area of research that we reviewed at the start of this section, rather than seeing multimethodology as an end in itself. It would be wise to keep in mind the sensible advice from Bryman (2007, p.8):

> *"The key issue is whether in a mixed methods project, the end product is more than the sum of the individual quantitative and qualitative parts."*

(Note that in the above quote Bryman takes a somewhat narrower view of multimethodology by equating it exclusively with a mix of quantitative and qualitative methods.) Bryman observes that social scientists who employ multimethodology in their research do not seem to dwell on ontological and epistemological questions and tend to take a pragmatic approach towards such issues. He notes, however, that there are potential problems in having to bridge the 'ontological divides' associated with different paradigms, to which the concept of triangulation does not necessarily provide a satisfactory answer. (We shall return to the importance of scientific paradigms for multimethodology research in Chapter 7.) Mingers & Brocklesby (1997) summarise the potential barriers to the successful application of multimethodology research in general, and multiparadigm research in particular, as follows:

- Cognitive – the problem of an individual researcher moving easily from one scientific paradigm to another.
- Cultural – the extent to which organisational and academic cultures militate against multi-paradigm work.
- Philosophical – 'paradigm incommensurability', which would occur if methodologies are thought to belong to different paradigms that are considered to be incompatible with each other.

In the MSU-DES paper, there were no issues relating to multimethodology. The only methodology used was that of discrete-event simulation modelling, which is usually regarded as belonging to hard OR. We did not apply any soft OR methodologies in the form of problem-structuring methods in preparation of the simulation modelling, as the general outline of the client's problem was already reasonably clear from similar studies done by OR practitioners elsewhere. The necessary context-specific details were investigated through interviews with the staff of the Unit and others and by direct observations of the Unit's activities by members of the research team. Such details were directly incorporated into the construction of the successive simulation models. The research team consisted of a senior academic researcher in overall charge of the projects, several student researchers doing the majority of the fieldwork, and the medical and senior nursing staff in charge of the Unit. There were no specific 'cultural' problems between the academic researchers on the one hand and the staff of the Unit on the other, as a collaborative relationship had been built up over a comparatively long period of time, and both sides were quite familiar with each other's way of working.

6.2 Multimethodology in the Maritime and Coastguard Agency studies

In the MCA papers, a multimethodology approach was developed over time. Given the highly aggregate nature of the annual Coastguard statistics that were initially available, we focused first on developing a Logistic Regression model in order to identify statistically the main explanatory factors affecting the annual proportion of lives saved in each Coastguard district. As our detailed literature review showed that this process had not been modelled before in a way that was helpful to us, in the construction of our regression model we had to rely mainly on suggestions in the report by Lord Donaldson (1999). Moreover, as we had little or no access to relevant stakeholders at that stage, there was no opportunity for us to employ any soft OR methodologies in the form of problem-structuring methods.

Once the main explanatory factors had been reasonably well established, at least at a macro (aggregate) level, our focus shifted from conventional regression analysis towards efficiency analysis based on the closely-related methodology of Stochastic Frontier Analysis (SFA). Given this new focus, it was then a logical step to compare the SFA results with those from another methodology for efficiency analysis, namely Data Envelopment Analysis (DEA). Both SFA and DEA are based on the idea of constructing a composite efficiency indicator for each DMU (in this case, each individual Coastguard coordination centre in each year). Whereas in other efficiency studies the input and output variables for a DEA model are often posited on more or less ad-hoc grounds, in the MCA case we took these variables directly from the earlier regression analysis, in order to make the results as comparable as possible.

There are, however, a number of important differences between SFA and DEA. One difference is that SFA is based on stochastic modelling and DEA on deterministic modelling. But in practice this difference is often not as great as might appear at first, as it is commonly acknowledged in DEA studies that the efficiency differentials found between DMU's should be interpreted with caution, because they may be attributable at least in part to stochastic factors. Another, perhaps more important, difference is that the variable weights in SFA are determined by estimated regression coefficients, whereas in DEA these weights are allowed to vary more or less freely (and may ultimately come to depend on the subjective evaluations of various stakeholders). But, as noted by Foreman-Peck (2002), this distinction is rather more important if we want to measure the composite values of different output combinations (as output weights may be strongly linked to the value judgements made by different stakeholders) rather than, as in the case of the MCA models, the effectiveness of different input combinations (about which different stakeholders may be expected to hold more neutral views). Like SFA, the way in which DEA was applied in the present case thus relied primarily on objective assessments of performance, rather than subjective evaluations, and, therefore, tended to belong to the realm of hard OR.

A much bigger methodological step was taken subsequently, when it was decided to develop a Bayesian Belief Net (BBN). However, the logic behind taking this step was again quite straightforward. The regression analysis had used aggregate data to estimate the significance of 'macro' factors (the average scale of incidents, the average workload of watchkeeping staff and the length of coast line monitored), in explaining the annual average success rate (in terms of the proportion of lives saved) for each Coastguard coordination centre. But it is not necessarily clear from regression analysis why macro factors such as the average scale of incidents or the average staff workload actually matter. Instead, it is obvious that there must be potentially many 'micro' factors at work in determining the risks involved in a SAR operation – at least some of which are recorded in the detailed incident records on which the annual Coastguard statistics are based. In short, the overall aim of this part of the analysis was not just to show the complex cause-and-effect relationships between the various risk factors at a micro level (that is, at the level of individual SAR incidents) but, thereby, also to gain a better understanding of how these factors tend to manifest themselves at the macro level (as addressed by the regression analysis). (To avoid potential confusion, we are not trying to argue here that the (later) development of the BBN was somehow used to justify the (earlier) selection of explanatory variables for the regression analysis. Rather, the BBN helps us to gain a better understanding of the meaning of these regression variables and the possible reasons for their estimated effects on the average success rate for each Coastguard coordination centre.)

Given that the methodology of BBN's had already been successfully applied to risk-analysis studies in some other areas (Sigurdsson *et al.*, 2001), it seemed reasonable to use this methodology in the present case as well. First, the annual Coastguard statistics were re-analysed using contingency tables, based on the explanatory variables already identified in the regression analysis. Such contingency tables provide estimates of conditional probabilities that are useful for the subsequent development of a BBN. Next, interviews were conducted with relevant experts from the MCA, supported by the detailed analysis of individual incident records held at MCA headquarters, in order to gather the primary data needed for the construction of the BBN. The MCA-BBN paper discusses the first stage of the BBN's construction (the development of its logical structure), while a paper explaining the second stage (the detailed quantification of the conditional probabilities in the BBN) is currently being revised for publication. This research is still ongoing. Indeed, further work is currently being done to compare the empirical evidence gathered from the analysis of the macro data to that gathered from the analysis of the expert interviews and the micro data.

Following Bryman (2007), we have already noted a key question relating to multimethodology research; namely, whether or not the end product is more than the sum of the individual parts. If we had treated the different methodologies – Logistic Regression Analysis, Stochastic Frontier Analysis, Data Envelopment Analysis, Bayesian Belief Nets – in relative isolation from each other, then multimethodology might not have brought any great benefits. But as we have attempted to demonstrate above, the results from the application of different methodologies in the MCA papers are directly complementary to each other. For instance, the construction of the (micro-level) BBN was directly informed by the results from the earlier (macro-level) regression and efficiency analysis and, in turn, the structure of the BBN sheds more light on the meaning of those earlier results.

As far as the potential problems for multimethodology research mentioned by Mingers & Brocklesby (1997) and Bryman (*op. cit.*) are concerned, the cognitive problem of using methodologies with which not everyone was very familiar was resolved by working as a team of researchers with complementary skills, some originally being trained in various forms of regression analysis, others in DEA, yet others in the construction of BBN's. However, the key requirement for a successful outcome was that all members of the research team – the membership of which tended to evolve over time – had to be willing to gain at least a basic familiarity with methodologies with which they had previously been unfamiliar. On the other hand, the cultural problem of switching between different research methodologies and, if appropriate, scientific paradigms is no real problem at all in a Department of Management Science that explicitly prides itself on its capability to accomplish this as a group of academic researchers.

Finally, the philosophical problem of paradigm incommensurability only became somewhat of an issue with the development of a BBN. The initial Logistic Regression Analysis and, as already explained, also the subsequent efficiency analysis (SFA and DEA) were primarily based on objective assessments of performance, belonging to the realm of hard OR. In terms of scientific paradigms, these methodologies all tend to embody the philosophical assumptions and principles of the paradigm of positivism, which combines a realist ontology (i.e. the assumption that an external reality exists, independent of our conceptions of it) with an epistemology based on the observation of empirical facts (with the ultimate aim of the verification or falsification of theoretical beliefs).

Unlike regression analysis – which is founded on a 'classical', objectivist approach to statistics – BBN's are founded on a 'Bayesian', subjectivist approach to statistics. While this does not necessarily mean that we have to abandon our idea of an external reality – indeed, many Bayesian statisticians might be extremely reluctant to do so, as this would negate a central element in the process by which 'prior' probabilities are to be transformed into 'posterior' probabilities – it does imply that the Bayesian approach explicitly allows for different beliefs about that reality between different people (expressed as different individuals holding different subjective probabilities of a particular event). Note that this last point is particularly relevant in the case of the MCA studies where – as we have explained in Section 3.2 about the political nature of decision making in the MCA – there still are wide differences in opinion between different groups of stakeholders about key issues; in particular, the impact of the closures of some coordination centres. In other words, seeing the methodology of BBN's as a straightforward embodiment of the paradigm of positivism (with its reliance on the undisputed observation of empirical facts) would appear to be rather inappropriate.

There were a number of reasons why any problem of paradigm incommensurability in the MCA studies ultimately seemed to be relatively minor. A key reason was that, as explained in detail in the MCA-BBN paper, we followed a well-established and rigorous process in conducting the interviews with the MCA experts that formed the main input into the construction of the BBN. This process was designed to minimise as much as possible the effects of various biases that are known to affect expert judgements in such situations. Moreover, we could exercise some level of empirical control through our statistical analysis of SAR incident data (the aggregate Coastguard statistics and subsequently also some of the individual incident records). Perhaps the main safeguard, however, was that the interviews with the MCA experts did not touch directly on controversial issues such as centre closures, but were concerned with the much more general question of what are the main risk factors determining the outcome of an individual SAR operation.

However, with hindsight, it might have been better if we had addressed the role of scientific paradigms more explicitly in our studies, particularly the problems associated with the application of an empiricist or positivist paradigm in this area of research. We shall, therefore, return to that point at greater length in the next chapter.

7 MODEL VALIDATION

7.1 Two popular scientific paradigms

As we have discussed, multimethodology may fail because of paradigm incommensurability, if methodologies are thought to belong to different paradigms that are considered to be incompatible with each other. If we are minded to accept the widely-quoted arguments by Burrell & Morgan (1979), this would appear to be a particular problem in the general area of social science (or, more particularly, organisation and management) research, of which our research on performance measurement in public service organisations would form but a specific example. According to the well-known framework proposed by Burrell & Morgan, social science researchers can choose between a number of co-existing but essentially incommensurable paradigms that can be distinguished along to two different dimensions: objectivity versus subjectivity, and regulation and stability versus radical change.

Burrell & Morgan's work has given rise to an extensive and ongoing debate in the academic literature on organisation and management research, not just about the specific dimensions of their framework (e.g. see Deetz, 1996), but also the extent to which the various paradigms really are incommensurable (e.g. see Gioia & Pitre, 1990; Schultz & Hatch, 1996; Tashakkori & Teddlie, 1998; Goles & Hirschheim, 2000; and Aram & Salipante, 2003). Nevertheless, there is a widespread acceptance – including among members of the wider community of OR practitioners (e.g. see Roy, 1993; and Pidd, 2003) – that there are crucial differences between a paradigm incorporating a realist ontology together with an empiricist or positivist epistemology at one possible end of the spectrum (akin to hard OR modelling), and a paradigm incorporating a non-realist ontology together with a constructivist or interpretivist epistemology at the other end (akin to soft OR modelling). For OR modelling, Mingers (2006b, p. 1378) summarises this point as follows.

> *"... [T]aking OR as an academic discipline, there has been no agreed and coherent underpinning philosophy, but simply several competing philosophies, the main ones, positivism and interpretivism having clearly incompatible assumptions."*

Each of these two different paradigms could be – and, indeed, often is – used to guide organisation and management research. But before addressing the question of which of these two widely-used paradigms – a realist ontology together with an empiricist/positivist epistemology (as appropriate to hard, mainly quantitative, research methodologies) versus a non-realist ontology together with a constructivist/ interpretivist epistemology (as appropriate to soft, mainly qualitative, research methodologies) – we should (have) select(ed), we should investigate whether they are indeed inherently incommensurable. One of the most fundamental questions in research philosophy relates to the nature of reality. Are we willing to adopt as a basic axiom either that an external reality exists independent of our conceptions of it (that is, a realist ontology) or, instead, that any notion of reality somehow cannot be divorced from our own and other people's thoughts about it (that is, a non-realist ontology)? A realist ontology may be taken to imply that some forms of knowledge are more valid than others (in the sense of corresponding more closely to the external reality that we, and others, are studying), whereas the latter statement becomes rather more problematic if we adopt a non-realist ontology. Hopefully, this very brief discussion is sufficient to convince the reader that philosophical questions about the nature of reality are so fundamental that any scientific paradigms incorporating a realist ontology are, in any normal sense of the word, incommensurable with any other paradigms incorporating a non-realist ontology. (This conclusion seems to match Morgan & Smircich's (1980) contention that the key dimension guiding the choice of research methodology in the social sciences should be 'objectivity' versus 'subjectivity', with 'reality as a concrete structure' at the extreme objectivist end of the spectrum, and 'reality as a projection of human imagination' at the extreme subjectivist end.)

Given their inherent incommensurability, which of the two popular paradigms mentioned above should we select to guide our own studies into performance measurement in public service organisations? First, let us consider the option of a realist ontology together with an empiricist/positivist epistemology. The key problems here are related to the nature of the epistemology – that is, the assumptions made by a researcher about how to gain valid knowledge. An empiricist/positivist epistemology, which is invariably combined with a realist ontology, implies that a researcher can gain valid knowledge through the careful observation of empirical facts that can be independently verified by other researchers and that can provide, through a process of verification or falsification of appropriate hypotheses, a better understanding of the external reality facing all researchers. However, in our specific area of research there are a number of strong arguments against an empiricist or positivist epistemology, as follows.

> First, different stakeholders tend to emphasise different performance dimensions. But note that this is not necessarily, in itself, a sufficient argument against an empiricist/positivist epistemology in this case. As Buchanan *et al.* (1998) argue, the notion of an objective reality faced by managerial decision makers can, and should be, carefully separated from the subjective preferences of these decision makers about aspects of that reality. In other words, different decision makers in a public service context could agree about the real nature of an organisational outcome as measured along its main performance dimensions, while possibly disagreeing about the desirability of that outcome.
>
> Second, performance measurements may also depend on the subjective judgements of different stakeholders (cf. Wisniewski & Stewart, 2004). This is a more serious objection as decision makers could in this case disagree about the real nature of an organisational outcome as measured along some key performance dimension – which makes it rather more difficult to talk about 'undisputed' empirical facts.

Third, in general, different stakeholders will have their own, potentially differing, perspectives on reality and will also contribute to shaping that reality in different ways (Danermark *et al.*, 2002). This is arguably the most serious objection as it calls into question the whole idea that external reality solely manifests itself through empirical facts that look substantially the same to all possible observers and the nature of which facts those observers cannot significantly alter themselves.

While objections to empiricism or positivism comparable the ones listed above are commonly raised in the wider area of organisation and management research, they are also frequently voiced in the specific research area of performance measurement in public service organisations. A typical example of the latter is the following quote from Williams (1993, p. 655).

> *"The empiricist can no longer be the arbiter of public morality, as he or she is all too often thought to need moral instruction. ... [P]erformance measurement may provide a false sense of objectivity when political constraints are ignored."*

But if the combination of a realist ontology and an empiricist/positivist epistemology is not acceptable, what about the alternative combination of a non-realist ontology and a constructivist/interpretivist epistemology? In that case, the key problems are related to the nature of the ontology rather than the epistemology. In our specific area of research there are a number of strong arguments against a non-realist ontology, as follows.

First, from a philosophical point of view, it would be hard to make much sense of the problem of unintended consequences that – as we have explained in Section 5.1 – appears to bedevil managerial attempts at performance improvement in many public service organisations (Bevan & Hood, 2006). If reality is merely what different people think it is, then how could reality end up acting in different ways from those intended? Obviously, there must be some mechanisms that act independently from people's thoughts and not all interpretations of reality are necessarily equally valid in explaining such mechanisms.

Second, from a political point of view, public accountability by the government can only operate in the context of independent scrutiny and well-informed debate (Bird *et al.*, 2005). But, again, if reality is merely what different people think it is, then how could one group of stakeholders (the wider public, including the electorate) ever hold an other group of stakeholders (the politicians overseeing the provision and management of public services) effectively to account?

Therefore, we may conclude that neither of these popular paradigms is, on its own, wholly suitable for our specific area of research. The first alternative incorporates an empiricist/positivist epistemology that tends to ignore the different perspectives – on the desirability of potential outcomes but also the probability of their occurrence as well as their essential nature – which different stakeholders in the public sector may hold and that, moreover, tends to deny the roles that these stakeholders may play in shaping potential outcomes. The second alternative incorporates a non-realist ontology that cannot cope very well with vital issues in public sector performance measurement such as the problem of unintended consequences or the need for public accountability.

7.2 How to deal with paradigm incommensurability

In Chapter 6 we discussed the benefits of taking a multimethodology approach to OR modelling. But we explained that multimethodology research, at least in its multiparadigm form, suffers from the problem of paradigm incommensurability. Given that we have now concluded that each of the two paradigms that we posited above incorporates some important aspects that are inappropriate for our specific area of research, and also given that these paradigms cannot be easily joined (with the aim of combining their strengths while masking their weaknesses) because of their inherent incommensurability, then what kind of approach should we adopt? Only applying methodologies within a single one of the popular paradigms (in terms of OR modelling, either hard OR or soft OR)? Or could we find a more creative solution to this conundrum?

From the relevant academic literature, there would appear to be three different ways of dealing with paradigm incommensurability.

1. Take a pragmatic approach and largely ignore the problem.
2. Apply a contingency approach to model validation, possibly combined with triangulation through reciprocal validation of the results from different methodologies.
3. Adopt a new philosophical position that can incorporate both hard, mainly quantitative, and soft, mainly qualitative, research methodologies – and thereby gaining a new perspective on multimethodology.

7.2.1 Taking a pragmatic approach

Taking a pragmatic approach seems to be the most common choice in organisation and management research (Tashakkori & Teddlie, 1998; Aram & Salipante, 2003; Morgan, 2007). For example, the 'mode 2' method of knowledge production (Gibbons *et al.*, 1994) has been championed among British researchers by Tranfield & Starkey (1998). Mode 2 research is to be conducted by 'trans-disciplinary' teams of researchers, engaging with both the world of theory and the world of practice – based on Kurt Lewin's (1951, p. 169) statement that "there is nothing so practical as a good theory". Likewise, Pragmatism has gained ground in OR modelling (Ormerod, 2006a). The report by the 'Commission on the Future Practice of OR' set up by the OR Society in the 1980's (Mitchell, 1986) already claimed that 'pragmatism rules' in OR modelling and little seems to have changed since then (Pidd, 2001; Munro & Mingers, 2002; Pidd, 2005).

We have already briefly referred to Pragmatism as a scientific paradigm in Section 4.2. Mingers (2000a, p.1258) summarises this philosophical position as follows.

> *"Pragmatism is a view about the purpose of science, that it is essentially a practical activity aimed at producing useful knowledge rather than understanding the true nature of the world."*

While Ormerod (2006b, p. 1373) explains his own motivation for taking a pragmatic approach in the following extended quote.

> *For me, action (the process of engagement, practice) comes first; interest in solving a problem is the motivating force; in acting, practitioners develop and deploy their craft skills (derived primarily from experience but also from colleagues, common-sense, intuition and rational argument – but unlikely to include a knowledge of philosophy); methods (only weakly linked to philosophy, if at all) may be called on to support action; reflection on action may or may not lead to philosophising, but it usually won't; philosophy may be a source of new ideas, but these ideas have to be validated in practice. Do they work? Are they useful?"*

Bryman (2007, p.17) argues that these kind of sentiments are widely-held among 'mixed methods' (i.e. multimethodology) researchers.

> *"Typically ... mixed methods researchers seem not to dwell on epistemological and ontological issues and exhibit a clear pragmatism in their work."*

A reasonably clear explanation of how the problem of paradigm incommensurability could be resolved by taking a pragmatic approach is given by Morgan (2007, p. 68).

> *"... [W]hat I am calling the pragmatist approach does not ignore the relevance of epistemology and other concepts from the philosophy of knowledge. It does, however, reject the top-down privileging of ontological assumptions in the metaphysical paradigm as simply too narrow an approach to issues in the philosophy of knowledge. ... The pragmatic approach that I am advocating would concentrate on methodology as an area that connects issues at the abstract level of epistemology and the mechanical levels of actual methods."*

In an accompanying diagram, Morgan places methodology at the centre, linking epistemology at the top and methods at the bottom – with ontology omitted altogether. As we have argued that the incommensurability of the two popular paradigms posited above is primarily at the ontological level, by neglecting ontological concerns – as Morgan advises – we can also ignore any problem of paradigms being incommensurable at the ontological level. (Cf. the advocacy by Johnson *et al.* (2007) of 'Pragmatism of the Middle' as an especially useful philosophy for multimethodology research, which seems to refer to a 'moderately weak' form of realism.) In short, according to this line of reasoning, taking a pragmatic approach means that we should largely ignore any problem of mixing realist and non-realist ontological perspectives. But that implies that we would be ignoring the issue that tends to give rise to paradigm incommensurability in the first place. It could be concluded that this way of bridging the 'ontological divide' is an intellectual cop-out. Moreover, it does insufficient justice to the kind of arguments that we advanced against a non-realist ontology in Section 7.1.

7.2.2 Applying a contingency approach to model validation

A different way of dealing with paradigm incommensurability, although in some sense connected to – if not part of – the pragmatic approach is to apply a contingency approach to model validation. It is now widely believed that there are no universal criteria for validating social science theories or models. In the more general social science literature, authors such as Lincoln & Guba (1985, and, with the names reversed, 1994), Sale & Brazil (2004), and Johnson *et al.* (2006) have suggested that different research methodologies require different validation criteria. In particular, they tend to claim that the use of hard, mainly quantitative, methodologies in social science research should be evaluated according to different criteria than the use of soft, mainly qualitative, methodologies. For hard, mainly quantitative, methodologies – which are normally applied in the context of a realist ontology – a set of (what we might call) 'classical' validation criteria (as outlined by Yin (2003), for instance) is generally taken to be appropriate:

1. Construct validity – establishing correct operational measures for the concepts being studied.
2. Internal validity (for explanatory or causal studies only, and not for descriptive or exploratory studies) – establishing a causal relationship, whereby certain conditions are shown to lead to other conditions, as distinguished from spurious relationships.
3. External validity – establishing the domain to which a study's findings can be generalised.
4. Reliability – demonstrating that the operations of a study (such as the data collection procedures) can be repeated, with the same results.

But Lincoln & Guba (*op. cit.*), in particular, have argued that the above criteria are inappropriate for soft, mainly qualitative, methodologies – where the ontology may be either non-realist or (in a pragmatic fashion) not precisely specified – and that they should, therefore, be replaced by a list including 'confirmability', 'credibility', 'transferability', and 'dependability'.

A similar debate has taken place over the last decade or so in the OR literature. For instance, Miser (1993, p. 212) defines validation as follows.

> *"Validation is the process by which scientists assure themselves and others that a theory or model is a description of the selected phenomena that is adequate for the uses to which it will be put."*

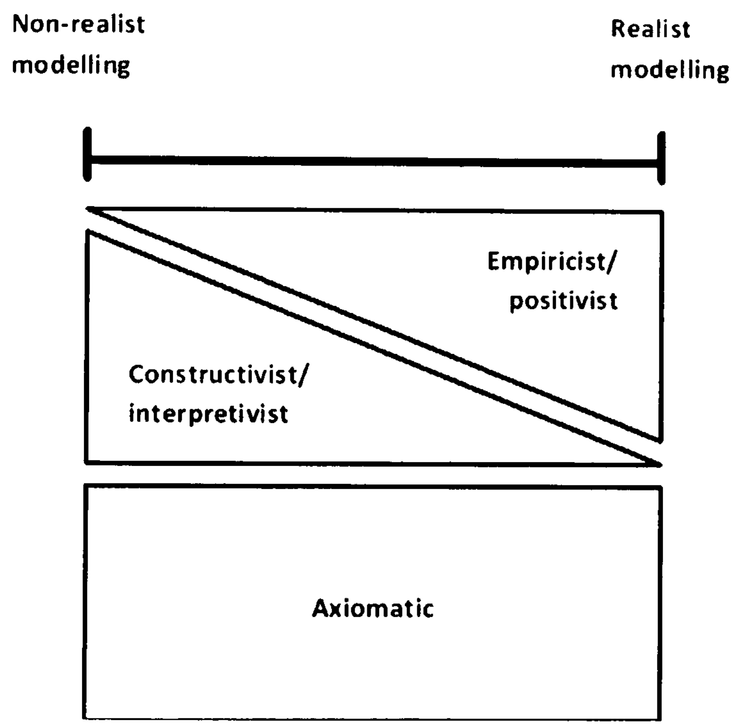
Miser (*ibid.*, p. 213) draws the following two inferences from his discussion of model validation in OR:

1. *"There are no universal criteria for validation. Rather the basis for judging the confidence that one should have in a model rests on the situation being modelled and the work that has been done, both in formulating the model and in comparing its consequences with reality."*
2. *"Any validity judgement is relative in at least two ways: with respect to the phenomena being modelled and the uses to which the model will be put. Thus a model may be used with adequate confidence in one situation but not in another."*

Miser's first conclusion – that there are no universal criteria for validation – is shared by Déry *et al.* (1993) and by other OR practitioners since. And his second conclusion – about the importance of 'usefulness' for an OR model – is also widely acknowledged. But Pidd (2003), drawing heavily upon Roy (1993), attempts to go beyond this by suggesting that – in OR modelling as in more general social science research – different methodologies should be evaluated according to different criteria, as illustrated Figure 3 overleaf.

Figure 3 is meant to show that the evaluation of a hard, mainly quantitative, OR model should be based on empiricist/positivist grounds, in which the model's correspondence with the objective external reality that it is meant to represent is of paramount concern. The evaluation of a soft, mainly qualitative, OR model, on the other hand, should be based on constructivist/empiricist grounds, in which the model's most valuable characteristic is its usefulness, both in terms of providing insights that were previously hidden or unknown ('acting as a set of keys to unlock doors through which to proceed') and in terms of moving an individual or group towards a commitment to action. However, in order for them to be truly scientific, both hard and soft OR models should also be judged according to axiomatic criteria – which in this context simply means that they should be consistent with their underlying frames of reference.

Figure 3. A contingency approach to the validation of OR models



(Adapted from Pidd, 2003, Fig. 11.4; based on Roy, 1993)

Pidd's contingency approach – which shares certain features with Oral & Kettani's (1993) contention that each different type of OR problem is associated with a different set of validation requirements – was, strictly speaking, developed for different types of single-methodology OR models (that is, either hard or soft OR, but not a mixture of both). However, it would seem to be no great step to presume that the same kind of contingency approach could be applied to multimethodology OR models; that is, models incorporating both hard, quantitative and soft, qualitative elements. In that case, the hard aspects of the model would be validated according to empiricist/positivist criteria and the soft aspects according to constructivist/empiricist criteria. This could then even lead to a process of triangulation by which the results from different – hard and soft – methodologies would be reciprocally validated. In the wider area of organisation and management research, this form of triangulation appears to be increasingly regarded as a key advantage of multimethodology research, if not its main purpose (Gioia & Pitre, 1990; Lewis & Grimes, 1999; Bryman, 2001; Johnson *et al.*, 2007).

Despite its growing popularity, the idea of triangulation is not unproblematic. Brannen (2005) demonstrates that the term can take a variety of meanings; but in social science research it is most commonly used to describe different research methods or methodologies corroborating each other. However, Brannen (*ibid.*, p. 12) goes on to argue that:

> *"...data collected from different methods cannot simply be added together to produce a unitary or rounded reality or truth. ... [I]f we move away from assuming that we are trying to arrive at a single reality, we need to understand how different accounts are arrived at and the purposes these accounts serve."*

Bryman (2007, p. 21) expresses a similar opinion, as follows.

> *"The metaphor of triangulation has sometimes hindered this process [of constructing a negotiated account] by concentrating on the degree to which findings are mutually reinforcing or irreconcilable. Mixed methods [i.e. multimethodology] research is not necessarily just an exercise in testing findings against each other. Instead, it is about forging an overall or negotiated account of the findings that brings together both components of the conversation or debate."*

In short, rather than trying to bridge the ontological divide between hard, mainly quantitative, and soft, mainly qualitative, methodologies by trying to establish a unified position on the nature of reality, these authors regard triangulation as a means towards 'forging a negotiated account' between different ontological positions. In other words, although the problem of mixing realist and non-realist ontological perspectives is no longer ignored – as suggested by Morgan (2007), for instance – we have still not arrived at a point where the problem of paradigm incommensurability can be said to have been fully resolved.

7.2.3 Adopting a new philosophical position and thereby gaining a new perspective on multimethodology

The final way of dealing with paradigm incommensurability is to adopt a new philosophical position that can incorporate both hard, mainly quantitative, and soft, mainly qualitative, research methodologies. Whether such an approach is viable depends crucially on the answer to a key question; namely, whether or not particular research methodologies are inextricably associated with specific ontological and/or epistemological axioms. To be more specific, are hard, mainly quantitative, methodologies inextricably connected to a realist ontology and an empiricist or positivist epistemology? Or are soft, mainly qualitative, methods inextricably linked to a non-realist ontology and a constructivist or interpretivist epistemology? Although there still seems to be a widespread belief in the more general social science that that is indeed the case (cf. the large-scale literature review conducted by Sale and Brazil, 2004), the correct answer to these questions would appear to be 'no' (cf. the arguments made by Halfpenny, 2005, as quoted in Brannen, 2005).

A philosophical position that would allow us to deal with our particular conundrum – namely, how best to cope with the apparent flaws in the two popular paradigms that we posited at the beginning of Section 7.2 and how to combine hard, mainly quantitative, and soft, mainly qualitative, research methodologies – would appear to be Critical Realism. This represents a comparatively recent development in the philosophy of knowledge. Critical Realism was first developed by the philosopher

Bhaskar (1978), who was strongly influenced by Harré's (1970) criticism of positivism. Since then, a substantive literature has grown around this topic, with the book by Danermark *et al.* (2002) as a particularly clear and useful explanation of its potential role in social science research. In OR, Critical Realism has found an ardent champion in Mingers (2000a and 2006a).

(Note that Flood & Jackson (1991)'s concept of "Total Systems Intervention" (TSI) is an alternative candidate for dealing with the issue of paradigm incommensurability in OR modelling. Although the TSI approach holds that different scientific paradigms are essentially incommensurable, it purports to provide a 'meta-paradigm' for choosing the most appropriate methodology for any particular OR problem. But as Mingers (2005) has pointed out, rather than fully resolving the problem of paradigm incommensurability, the TSI approach is essentially another contingency approach that relies on the selection of a single methodology depending on the particular context.)

As a philosophical position, Critical Realism is founded on a realist ontology that can nevertheless be combined with a constructivist or interpretivist type of epistemology. In fact, Critical Realism gives prominence to ontology rather than epistemology. It rests on the basic assumption that objects in reality possess causal powers, i.e. 'generative mechanisms'. Reality is differentiated and stratified.

First, reality is differentiated, because three different domains of reality can be distinguished:

- The 'real' domain, where the generative mechanisms are to be found. These mechanisms exist irrespectively of whether they produce an event or not.
- The 'actual' domain, where events actually produced by the mechanisms are to be found.
- The 'empirical' domain, where those events that are experienced are to be found.

In other words, generative mechanisms do not always produce events, and any events actually produced are not always experienced. Therefore, the reality that scientists study is larger than just the empirical domain.

Second, reality is also stratified in that generative mechanisms belong to different 'strata' of reality. One can talk, for instance, of chemical, biological, physical, psychological and social strata. (From an OR perspective, Mingers (2003) distinguishes between the material, personal and social dimensions of the world.) Although the exact classification of these strata is still a matter for debate, the important point is that new mechanisms are continually created in their respective strata. Such newly-created mechanisms are termed 'emergent' powers. In the specific context of organisation and management studies, Fleetwood & Ackroyd (2000, p. 7) note that:

> *"organisations are structures that are reproduced by the participants in them, but they have emergent properties that bind participants into a particular pattern of relationships."*

In other words, an organisation (or any other kind of social structure) is not just the result of actions of individual agents; instead, the organisation has a real existence (comparable to that of material objects) which in itself enables or constrains the actions of individual agents.

Critical Realism thus assumes that an external reality exists, independent of our conceptions of it. This reality is the 'intransitive' object of science. The overarching purpose of science is to come as 'close' to this reality as possible. However, our knowledge of reality (consisting of our theories, models etc.) is always fallible and more or less 'truthlike'. As such knowledge is socially determined and changeable, it constitutes a 'transitive' object. One of the main reasons why scientific knowledge is inherently fallible is that reality consists of many different objects, which – because of their distinct structures – also possess different mechanisms or powers. This means that, in general, many different mechanisms are operating at the same time and that the relation between the generative mechanisms and their empirical effects is contingent. In other words, the events that we can observe constitute complex combinations of the influences of different mechanisms – some mechanisms reinforcing each other, other counteracting. This implies that causal laws (linking empirical events to the underlying mechanisms) must be analysed as 'tendencies' – and definitely not as the universal empirical regularities that positivism would seek to discover.

In general, the higher the stratum of reality, the more emergent powers there are and thus the greater the multitude of mechanisms operating singly or in combination. Therefore, all sciences that study the higher – in particular, the social – strata of reality, are practised in 'open' systems, in which it is more or less impossible to observe a particular generative mechanism in isolation and independently of other mechanisms. This goes for the social sciences in general, and obviously also for organisation and management studies in particular.

There is then a key difference between the natural or physical sciences on the one hand, and the social sciences on the other. In the natural or physical sciences, studying the lower strata of reality, it is generally possible to performing scientific experiments in which closed systems are created artificially. In such an experiment, a particular mechanism is observed in isolation, so that the researcher can draw a direct link between the mechanism operating in the real domain and its observed effects in the empirical domain. This is the basis for empiricism/positivism as a philosophy of knowledge. But in the social sciences, such experiments are generally not feasible. In the latter context, Danermark *et al.* (2002, p. 200) note the following:

> *"First, society [and any organisation within society] is made up of thinking and reflective human beings. They are capable of continually changing the social reality. Therefore, one might say that the reality social scientists [including organisation and management scientists] study is socially produced. Second – and what is more decisive for methodological issues – we*

study other people's interpretations of the social world. Our object of study is thus socially defined. We interpret the interpretations of other people."

Because organisations have emergent powers in the form of human intentionality, reflection, language and a capacity for self-change, their study tends to be pursued in an open system – for which reason, a positivist approach is generally unsuitable for organisation and management studies. Instead, Critical Realism aims to combine ontological realism with a kind of epistemological relativism or interpretivism (cf. Archer *et al.*, 1998; Kwan & Tsang, 2001). However, although all knowledge is socially produced and indeed socially defined, Critical Realism does not accept that all knowledge is therefore equally valuable. We should distinguish between epistemological relativism and judgemental relativism. The former idea means that our interpretation of (other people's interpretations of) reality is historically determined and fallible, as has been argued already. But the latter would imply that there are no grounds for deciding when one kind of knowledge should be preferred to another – but that idea is decisively rejected by Critical Realism.

Critical Realism thus combines a realist ontology with an epistemology that is akin to constructivism or interpretivism. Therefore, its adoption would enable us to avoid the problems associated with the application of either an empiricist or positivist epistemology or a non-realist ontology to our specific research area of performance measurement in public service organisations. Moreover, Critical Realism contains the idea of 'Critical Methodological Pluralism' (Danermark *et al.*, 2002), through which we could satisfactorily resolve any problem of combining hard, mainly quantitative, and soft, mainly qualitative, research methodologies in our studies. (As long as we are able to treat all these different methodologies as consistent with a realist perspective on ontology.)

Critical Methodological Pluralism is simply the label given by Danermark *et al.* to the practical combination of different methodologies in an empirical study. Danermark *et al.* (*ibid.*, p. 162-163) explains the role of Critical Realism in enabling methodological pluralism as follows.

"Critical Realism is a metatheory, which enables us to understand the importance of methodologies in a partly new way. That is also the significant difference between our view and the pragmatic one. The decisive question is how different methodologies can convey knowledge about generative mechanisms. As we have seen, mechanisms are regarded as tendencies which can be reinforced, modified or suppressed in a complex interaction with other mechanisms in an open system. The result may be that they cannot always manifest themselves empirically. In addition, the [human] motive for action is regarded as a causal mechanism beside others, which makes the traditional division between a quantitative and explanatory methodology on the one hand, and a qualitative and understanding methodology on the other hand, limiting and misleading."

Instead of a false (or at least unhelpful) dichotomy between soft, mainly qualitative, and hard, mainly quantitative, methodologies, Danermark *et al.* (*ibid.*, p. 163) follow Sayer (1992) in distinguishing between 'intensive' and 'extensive' empirical procedures.

"[T]he intensive empirical procedure contains substantial elements of data collecting and analyses of a qualitative kind. The extensive procedure has to do with quantitative data collecting and statistical analysis. ... What we are discussing here [are] complementary empirical procedures and their being part of a greater whole, namely the research process guided by a Critical Realist ontology."

In short, Critical Methodological Pluralism implies a multimethodology research design combining extensive and intensive empirical practices. According to Danermark *et al.* (*ibid.*, p. 17 6):

"[it] is critical, first in the sense that it opposes an unreflecting employment of methods; the choice must be grounded in metatheoretical consideration.

Second, the label alludes to the methodological approach which is founded on the Critical Realist conception of ontology."

At this point in the discussion, it is important to stress that we have not explicitly tried to take a Critical Realist position in any of our four papers currently under consideration. Instead, we tended to take the pragmatic view of largely neglecting ontological concerns. As already explained in Section 6.2, in our own studies any problem of paradigm incommensurability ultimately seemed to be relatively minor. Without necessarily saying this in so many words, we always stuck to a realist perspective on ontology and tended link this to a generally empiricist or positivist stance on epistemology. Our realisation of the problems associated with the application of an empiricist or positivist epistemology in our specific area of research only developed as we were working on the construction of a BBN for the MCA case. With hindsight, that could have been an appropriate stage in our research to explore the Critical Realist position more seriously – nevertheless, we intend to do just that in our ongoing research on performance measurement in public service organisations.

7.3 Validation criteria for OR modelling

We have already discussed certain aspects of model validation in Section 7.2.2. In particular, we noted two widely-accepted features of validation of OR models; namely, the lack of universal criteria for validation and the importance of usefulness as valuable model characteristic. We also examined the contingency approach to model validation – which would suggest different validation criteria for soft, mainly qualitative, OR models, as opposed to hard, mainly quantitative, ones – and concluded that it could not provide a fully satisfactory answer to the problem of paradigm incommensurability. Rather than taking a contingency approach, it would seem preferable to find a set of, suitably generic, validation criteria that could be applied effectively to a relatively wide range of different OR models. Adopting a Critical Realist position will enable us to achieve this goal in a way that fits in with our own basic tendency as OR practitioners to stick to a realist perspective on ontology.

There are potentially numerous criteria for validating theories or models, including pragmatic ones such as usefulness (Pidd, 2003). For example, Smith (1993) describes how researchers involved in the evaluation of the 'Expected Utility' model mentioned numerous alternative validation criteria apart from the classical ones (such as axiomatic validity), including usefulness, comfort, the degree to which the model aids communication, whether it provides technological validity (an audit trail), do people come back and use it again, does it enhance decision making, etc. But in Critical Realism the key quality criterion for a theory or model is its 'explanatory power'. Danermark *et al.* (2002, p. 148) express the general research logic as follows.

> *"[Social science theories] are evaluated with respect to such criteria as explanatory power, the ability of the theories to conceptualise fundamental social mechanisms and integrate central concepts from other theories, whether they are creative or not, and whether they are logically consistent."*

In his book on Critical Realism, Lawson (1997, p. 243) makes the point more forcefully.

> *"Although the criteria for accepting knowledge claims as potentially true may include pragmatist and coherent ones, if amongst others, the commitment to realism entails that such criteria must also, and primarily, turn on the power of some theoretical claim to illuminate a (relatively) mind-independent reality. Truth judgements, on a realist view, will incorporate descriptive and evidential components <u>alongside</u> any pragmatist and coherentist considerations."* (Emphasis in original.)

In words more familiar to OR practitioners, Critical Realism strongly favours open-box (or white-box) validation – at least in a form in which the three different domains of reality are clearly distinguished and in which the causal relationships 'within the box', as well as their underlying assumptions, are themselves opened up to empirical scrutiny. Black-box validation, at least in its purest form of solely considering a model's predictive power with no regard to its internal structure, would be meaningful only in conditions of strict experimental control. (Note , however, that – just as internal and external validity tests are not mutually exclusive but are both important for evaluating the quality of a theory or model – black-box and white-box validation may be combined in practice into some form of 'grey-box' validation.)

We propose that those OR models that are founded – like ours – on a realist ontology should be validated according to two groups of validation criteria. The group of primary criteria would consist of the four classical validation criteria outlined by Yin (2003) and many other social science researchers (as explained in Section 7.2.2): construct validity, internal validity, external validity, and reliability. The group of secondary criteria would consist of a variety of pragmatic criteria to evaluate the usefulness of the model under consideration.

The group of primary criteria is based on a combination of realist and axiomatic perspectives on model validation. More specifically, the four classical validation criteria should be interpreted in light of the Critical Realist position, so as to make them equally applicable to hard, mainly quantitative (i.e. 'extensive') and soft, mainly qualitative (i.e. 'intensive') methodologies – as follows.

First, construct validity refers to establishing the correct operational measures for the concepts being studied. In a Critical Realist approach to OR modelling, concepts take the form of conceptual abstractions relating not just to the 'empirical' domain of reality, but also to the 'actual' domain and the 'real' domain. In other words, the question of construct validity goes well beyond quantitative measurement issues in the empirical domain and is equally (if not more so) relevant to the definition of qualitative constructs in the real domain (such as the generative mechanisms that are the ultimate focus of Critical Realist theorising).

Second, internal validity refers to establishing a causal relationship, whereby certain conditions are shown to lead to other conditions, as distinguished from spurious relationships. The main aim of the Critical Realist position is to gain knowledge about causal mechanisms linking the empirical, actual and real domains (but analysed as tendencies, not universal empirical regularities). Interpreted in this way, internal validity is the key criterion for evaluating an OR model, whatever the (mix of) research methodologies applied. In fact, as we have seen, Critical Realism advocates the application of complementary – hard/extensive and soft/intensive – methodologies (under the label of Critical Methodological Pluralism) to gain such knowledge.

Third, external validity refers to establishing the domain to which a study's findings can be generalised. According to the Critical Realist position, external validation is not based on checking whether empirical phenomena predicted by the model are generally occurring. Instead, generalisation is interpreted as using 'retroductive' inference (Danermark *et al.*, 2002) to find fundamental structures and causal mechanisms operating in the real domain, as well as establishing the region of time-space for which these mechanisms are actualised in the form of 'demi-regularities' (Lawson, 1997). Again, both extensive and intensive methodologies should be used for this task. Another requirement explicitly imposed by Critical Realism (but not always expressed as forcefully by other approaches) is to seek out competing models and then to try eliminate as many of these as possible on empirical grounds. A particularly strong model is one that could account for all of the empirical phenomena explained by rival models but, in addition, could explain some important phenomena that other models leave unexplained.

Fourth, reliability refers to demonstrating that the operations of a study can be repeated, with the same results. As explained by Tsang & Kwan (1999), although the general infeasibility of artificially creating closed systems in social science research – including organisation and management studies and, therefore, also OR modelling – makes replication difficult, it is not impossible in principle. In fact, Tsang & Kwan refer to six different types of replication that could be tried, each of which would apply to both extensive and intensive methodologies.

The group of secondary criteria is based on a combination of constructivist, instrumentalist and axiomatic perspectives on model validation. The main focus of these criteria is the usefulness of the OR model. However, usefulness can be interpreted in many different ways – as demonstrated by Oral & Kettani (1993), Smith (1993), and many other OR practitioners since. So, as the usefulness of an OR model cannot be divorced from the uses to which it may be put (Miser, 1993; Pidd, 2003), it is impossible to give a definitive and exhaustive list of headings for this group. But we would argue that all of these secondary criteria should be understood on the basis of Fleetwood & Ackroyd's (2000) contention that Critical Realism as a metatheory should guide not just the theory-related aspects of OR modelling, but also the practical use of such models for policy development and managerial intervention.

In summary, if one is willing to accept Critical Realism as a metatheory to guide OR modelling, then there is no need to resort to a contingency approach to model validation – according to which one would use different validation criteria for soft, mainly qualitative, OR models, as opposed to hard, mainly quantitative, ones. Instead, the two groups of validation criteria explained above could be applied to a wide range of different OR models – whether incorporating hard/extensive or soft/intensive methodologies or a mixture of both – as long as one is willing to accept a realist perceptive on ontology.

7.4 The validation of the OR models

Although we had not formulated the validation criteria for our own models in the precise manner explained above, both in the MCA studies and in the MSU-DES paper we did, in fact, apply a range of criteria of both a classical (primary) and pragmatic (secondary) nature. Overall, as already explained, we stuck to a realist ontology and we tended to look for open-box rather than black-box validation.

In the MCA papers, our main focus has been on the classical validation criteria of construct validity, internal and external validity, and reliability. As we started with multiple regression analysis, we used standard econometric procedures for hypothesis testing in order to identify the main explanatory variables influencing the performance of the Coastguard service at the macro (aggregate) level. (But see Mingers (2006a) for a critique of this kind of 'null-hypothesis significance testing'.) We also conducted some form of forecasting test, taking care to use different data sets for building the model (annual data for 1995-1998) and testing it (annual data for 1999). The results from our Logistic Regression model were then compared to those obtained through Stochastic Frontier Analysis (SFA); and, in turn, the composite efficiency indicators from the SFA model were compared with the ones we found through the use of Data Envelopment Analysis (DEA). These different kinds of validity checks represented some form of grey-box validation in which we were mainly concerned with construct validity, external validity and reliability, but where the internal structure of the model was only considered at the macro level. With the development of the micro-level Bayesian Belief Net (BBN) we moved much further towards open-box validation, although the reliability of the research process remained an issue of particular importance from a validation perspective. Although we have described in the MCA-BBN paper our initial efforts at triangulating our empirical findings from the macro and micro levels of analysis, this remains an ongoing challenge.

In terms of usefulness criteria, the key purpose of our MCA papers was – as we have explained in Chapter 4 – to expose managerial decisions about the organisation of the

Coastguard service to public accountability, rather than to aid managerial decision making in any direct sense (at least until now). From a pragmatic point of view, therefore, we can argue that the usefulness of our MCA models was best served by trying to apply the classical validation criteria as effectively as possible – but with a particular emphasis on making the internal structure of our models more transparent to non-specialists in OR, as we were attempting to do through the development of the BBN.

In the MSU-DES paper, we have tended to take care of the classical validation criteria by applying Robinson's (2004) guidelines for the validation of Discrete Event Simulation models. Robinson distinguishes between no less than seven different forms of validation, all of which are concerned with ensuring that important aspects of the model are 'sufficiently accurate for the purpose at hand': conceptual model validation; data validation; open-box validation; black-box validation; experimentation validation; solution validation; and verification. Of this list, the two forms of validation that turned out to be most problematic in our MSU study were black-box validation and solution validation. The internal structure of each one of the sequence of simulation models was extensively checked, in close cooperation with our clients, in the course of an ongoing process of open-box validation. As regards black box validation, however, our models were not always able to predict accurately the real system's response to some types of demand stimuli. As we have explained in Sections 2.5 and 3.6, we were not always accurately informed of temporary rises in staff availability in order to deal with particularly heavy patient loads in the Unit. As a result, the capacity of the Unit to cope with such patient numbers appeared to be rather greater than we would have predicted from our simulation models. While from a validation perspective it would be helpful if the real system could be operated in a reasonably stable fashion, this was not always acceptable from a managerial point of view.

The apparent difficulty of solution validation, on the other hand, is closely connected to our growing realisation in the course of the MSU study that the aim of the modelling process was not so much to provide ready-made 'solutions' to the clients' problems, but to enhance their understanding of the main performance drivers of the Unit. While the results from simulated experiments can be used to make recommendations for system changes, only some of these recommended changes may eventually be implemented – and, if so, not necessarily in their original form. As explained in Chapter 4, a key purpose of our MSU study was, therefore, to facilitate organisational learning and thus to help our clients make a case for further investments in the service. This purpose should constitute the main basis for any usefulness criteria for our sequence of simulation models. Figure 7 (p. 170) in the MSU-DES paper is meant to summarise our views on this issue.

8 FINAL REFLECTIONS ON THE FOUR PAPERS UNDER CONSIDERATION

8.1 Structured comparison between the four papers

To start the final reflections on our four papers under consideration, we shall compare them along the following three aspects:

(1) The subject matter of the study and the nature of outcome measures
(2) The purpose of the study
(3) The kind of OR models applied and the method of validation

The subject matter of all four papers is similar. The three MCA papers deal with the coordination of Search And Rescue operations around the UK coast by the Maritime and Coastguard Agency; and the MSU paper is concerned with the provision of health care by a local Musculo-Skeletal Unit operating across two Glasgow NHS hospitals. Both are specific examples of UK public service organisations. But the studies of the two organisations differ in some important respects; principally, in terms of:

- the precise nature of the public services that we considered – emergency services in the case of the MCA studies as opposed to health services provided mainly on an elective basis in the case of the MSU study;
- the locus of organisational decision making on which we focused – the national management of the MCA versus the senior members of the medical and nursing staff of the local MSU;
- the nature of the outcome measures that we used – the proportion of lives saved as a key measure of service outcome for the MCA as an emergency service as compared to various output (rather than true outcome) measures, such as patient numbers treated and waiting times, for the MSU as a health care provider.

With respect to the last point in particular, for performance measurement to be truly effective it would always be preferable to use the most 'fair' and comprehensive measures of the outcome of a public service possible. For the MCA studies, we were able to construct a measure of service outcome in the form of the proportion of lives saved from all those involved in life-threatening incidents. But this measure is neither entirely 'fair' to the service (in particular, we did not have access to the detailed data needed to distinguish between 'genuine' maritime accidents and incidents involving suicide or criminal acts) nor fully comprehensive (in particular, our measure discounted all of the MCA's efforts to deal with non-life-threatening incidents by its very nature). For the MSU study, a true measure of service outcome would relate to actual improvements in the health status of eligible patients as a (more or less) direct result of providing the service. But, as is all too common in simulation models of the health service, we simply lacked the data to construct an outcome measure of that nature, forcing us to focus on a range of output measures, instead. But note that many stakeholders in the health service regard an output measure such as patient waiting times to be a highly important performance indicator in its own right!

Despite the various differences listed above, both the MCA and the MSU were confronted with the usual list of problems and challenges for performance measurement in public service organisations, as demonstrated in Chapter 3 above. In particular, disputes about the proper trade-off between service productivity and service outcomes, although not necessarily couched in these terms, played a central role in both the MCA studies and the MSU study. In the MCA study, the debate between the various stakeholders crystallised around the question whether the closure of a Coastguard coordination centre would result in a trade-off between staff experience (associated with higher labour productivity) and local knowledge (associated with worse service outcomes). In the MSU study, the key point at issue was how best to resolve the expected trade-off between patient numbers (associated with higher labour productivity) and waiting times (associated with worse service outcomes). In both the MCA and the MSU case, these disputes had strong political overtones, as predicted more generally by Williams (2003). It is arguable, at least, that such fundamental similarities in the types of issues faced tend to outweigh the differences in the nature of services, the locus of organisational decision making or the nature of service outcome measures across the studies.

As explained in some detail in Chapter 4 above, the overall purpose of the MCA studies differed in some key respects from that of the MSU study. A very important aim for all the MCA papers was to make a scientific contribution to public accountability of Ministers (and senior civil servants) for their stewardship of the public service in question. Establishing 'what works' in promoting stated objectives – and thereby laying the scientific basis for performance improvement – was a subsidiary aim, albeit of ever growing importance. For the MSU study, on the other hand, the principal aim was to find out what works well and where performance improvements can be made. We developed a much closer relationship with our clients in the latter study, as befitted our desire to facilitate organisational learning and thereby to aid managerial decision making.

The OR models developed in the MCA studies tend to correspond to Mitchell's (1993) first type of model: a statement of the beliefs held for securing better understanding; or, in other words, the models fit in well with the idea of OR as a science to aid theory development. In terms of the different categories of OR modelling approaches that we developed in Section 5.3 above, the DEA and SFA models fit category 4 and the BBN model category 3. The models in both of these categories could be termed "models for evaluating effects of changes in systems" (Sanderson & Gruen, 2006). To put it simply, the DEA and SFA models were intended to help us to uncover significant differences in performance and the BBN model was meant to give us a better understanding of the cause-and-effect relationships driving these performance differentials. As discussed in Section 7.4, the validation of the MCA models focused on the classical validation criteria of construct validity, internal and external validity, and reliability. However, properly triangulating our empirical findings from the macro (SFA and DEA) and micro (BBN) levels of analysis remains an ongoing challenge. Since, as explained in

Section 4.3, the MCA management seemed to be somewhat suspicious of the 'real' motive for our research at first, we did not assess the usefulness of the MCA models primarily in terms of their acceptability to them; instead, we were rather more concerned with how our models contributed to the promotion of public accountability.

The sequence of DES models developed for the MSU study, on the other hand, tends to correspond to Mitchell's (*op. cit.*) second type of model: a device for helping to select suitable problem-solving actions; or, in other words, these simulation models fit in well with the idea of OR as a science to aid problem solving. The MSU models fit category 3 in our list of OR modelling approaches; that is, they are "models for evaluating effects of changes in systems" – like the MCA models. As explained in Section 7.4, the DES models were subject to seven different forms of validation (following Robinson, 2004), of which black-box and solution validation turned out to be most problematic. Because the development of the MSU models was intended to facilitate organisational learning on the part of our clients, the usefulness of the MSU models was mainly judged in that light.

In conclusion, while there are similarities between the four papers in terms of the subject matter of the studies and the types of issues faced, there are also some differences with respect to the purpose of the studies and, along with this, the particular OR models applied. The promotion of public accountability by the government for the performance of a vital public service was a primary aim of the MCA studies; and the application of Frontier Analysis techniques constitutes a useful set of tools in pursuit of that aim. In contrast, the MSU study was principally aimed at enabling performance improvements to be made by local service staff; and Discrete Event Simulation models can be used to conduct simulated experiments that may form a basis for future problem-solving actions. But all of our OR models – across the four papers – shared a common characteristic in being particularly suited to evaluating the effects of changes in systems.

8.2 Final reflections on the four papers in relation to theoretical and empirical studies conducted by other authors

We shall now discuss each of the four papers separately, in relation to comparable theoretical and empirical studies conducted by other authors.

8.2.1 Final reflections on the MCA-LRA paper

First of all, we shall consider the MCA-LRA paper; in particular, the application of the (Binary) Logistic Regression Analysis model. (NB: In this thesis we have tended to refer to this model as the 'LRA' model, although it was originally called the 'BLR' model. This change in label is not meant to have any special significance.)

As already explained in Sections 4.3 and 5.5 above, our initial interest in the performance of Coastguard coordination centres had been raised by newspaper reports about the proposed closures of some centres. As we began our studies, we perceived a need to investigate what causes performance variations between individual Coastguard coordination centres and to what extent the Coastguard service was becoming more effective, or less, in saving lives over time. More particularly, we wanted to subject the key trade-off identified by Lord Donaldson (1999) – between the level of staff experience and their level of local knowledge – to scientific analysis based on a formal modelling approach.

8.2.1.1 Brief comparison with relevant studies conducted by other authors

There are number of strands in the academic literature that are potentially relevant to the MCA issue, including papers on the management of SAR operations and a somewhat separate set of papers on Maritime Risk. A good example of the first strand of literature is the paper by Abi-Zeid and Frost on 'SARPlan' (Abi-Zeid & Frost, 2005), which is also referred to in the MCA-LRA paper itself. Abi-Zeid and Frost provide the following brief explanation of SARPlan (Abi-Zeid & Frost, *op. cit.*, p. 630):

"SARPlan [is] geographic decision support system designed to assist the Canadian Forces in the optimal planning of search missions for missing aircraft. Its primary purpose is to ensure that the available search resources are deployed in a way that will maximise the mission's probability of success."

Although not primarily developed for that purpose, SARPlan could also be used in principle to support the coordination of maritime SAR operations. However, SARPlan is not designed to model the full range of variables that might influence the operational effectiveness of SAR operations in a maritime context. Moreover, a well-designed decision support system such as SARPlan complements both the amount of experience that members of staff have gained in the performance of their duties and their level of local knowledge; and it potentially ameliorates the terms of Lord Donaldson's trade-off between staff experience and local knowledge and, more widely, the trade-off between service productivity and service outcomes. But SARPlan does not assist us in modelling the terms of such trade-offs directly, as we would like to do in our MCA studies.

A second strand of academic literature explicitly addresses the problem of maritime risk. As many of these papers have already been reviewed in the MCA-BBN paper, we shall limit ourselves here to a very brief overview. This second strand includes, *inter alia*, a review paper on risk assessment in maritime transportation by Guedes Soares and Texeira (2001), a series of papers on the 'Policy for Sea Shipping Safety (POLSSS)' project – e.g. the paper by Walker (2000); a series of papers written by researchers from Liverpool John Moores University – for example, the paper by Eleye-Datubo *et al.* (2006); and a series of papers written by researchers from The George Washington University, Virginia Commonwealth University and Rensselaer Polytechnic Institute – for example, the paper by Merrick and van Dorp (2006). However, all of these papers focus on the organisational and situational risk factors leading to maritime accidents and the resulting potential for risk reduction. That is, they do not model the variables that contribute to a successful maritime SAR operation, once a potentially dangerous – more particularly, life-threatening – situation has actually developed. Therefore, they are not directly relevant to the specific issues addressed by our MCA studies.

8.2.1.2 Aspects of OR modelling

Given that our preferred measure of service outcome for the MCA studies is binary in nature – as explained in Section 5.5 above, the life of a person involved in a life-threatening incident is either saved or lost – our first set of OR models was based on (Binary) Logistic Regression Analysis (LRA). We applied this form of multiple regression analysis to aggregate Coastguard Statistics for the years 1995-98, which constitute a set of 'panel data'. Panel data are data that vary both across space and time. From an econometric perspective, if the number of cross-sectional units is relatively large (in our case, 21 different Coastguard coordination centres) and the number of time periods over which these units are observed is relatively small (in our case, four consecutive years), then there are two main ways of constructing a regression model based on panel data. Put simply, the choice is between a "fixed effects model" and a "random effects model" (Kennedy, 2008; see also Coelli *et al.*, 2005).

8.2.1.2.1 Developing and testing a fixed effects model

As noted in the MCA-LRA paper, the fixed effects model involves the assumption that the intercept in the multiple regression model may vary across the cross-sectional units (i.e. the 21 Coastguard coordination centres) and/or the time periods (i.e. the four years). An appropriately large number of dummy variables must then be added to the LRA model in order to measure such shifts in the estimated regression line between different coordination centres and/or years. (In the MCA case, a maximum of (21 – 2 =) 19 dummy variables were added for the different coordination centres, while none of the dummy variables for the different years turned out to be statistically significant.) As Kennedy (*op. cit.*) explains, the use of the fixed effects model is reasonable if the data set exhausts the population. But this is indeed the case for the MCA studies, because the annual Coastguard statistics cover the whole population of coordination centres.

Although quite reasonable from the point of view of econometric theory, the inclusion of up to 19 dummy variables in the LRA model presented a practical problem in the form of multicollinearity between the three main explanatory variables (measuring staff workload, scale of incidents and length of coastline, respectively) and at least some of the dummy variables – as explained in more detail in the MCA-LRA paper. Multicollinearity tends to lead to imprecise estimates of the individual regression coefficients – that is, the regression coefficients of the individual explanatory variables tend to have relatively large standard errors.

But apart from that particular problem, the overall fit of the LRA model was also not very impressive. The Hosmer-Lemeshow test represents one of the most widely-used goodness-of-fit tests for a Binary Logistic Regression Analysis model (Hosmer & Lemeshow, 2000). (The Hosmer-Lemeshow test is preferred over alternative tests, such as the Pearson test or the deviance test, when at least some of the explanatory variables included in the model are of a continuous – rather than discrete – nature, as is the case in our LRA model (Xie *et al.*, 2008).) The Hosmer-Lemeshow test involves partitioning the observations as evenly as possible into ten groups according to their predicted probabilities. The test statistic is then based on the sum of the squared differences, suitably scaled, between the observed number of cases in each group and the expected number of cases. (Roughly speaking, the smaller these differences, the better the fit.) The Hosmer-Lemeshow test statistic has a Chi-Square distribution with (10 – 2 =) 8 degrees of freedom. If the corresponding p-value is less than one's accepted α-level (usually 0.05), then the test would reject the null hypothesis of an adequate fit.

In the case of our LRA model, the value of the Hosmer-Lemeshow test statistic was strongly affected by how many (and also which) dummy variables were included. (This variation in the precise value of the test statistics is, again, a reflection of the multicollinearity problem mentioned above.) As reported in the MCA-LRA paper, when the maximum number (i.e. 19) of dummy variables was added for the different coordination centres, the Hosmer-Lemeshow goodness-of-fit test had a p-value of 0.123, which is only marginally in excess of the usual α-level of 0.05. In other words, although the test did not reject the null hypothesis of an adequate fit, its relatively low p-value did not constitute evidence for a particularly good fit. (Note that somewhat higher p-values could be achieved by judiciously excluding some of the statistically less significant dummy variables from the LRA model, but the fit never became very good.)

8.2.1.2.2 Conducting a prediction test

This impression of a regression model with an adequate, but not impressive, fit was strengthened by the results from the prediction test that we conducted on the LRA model. As explained, the LRA model had been estimated using aggregate Coastguard statistics for the years 1995-98. Given our subsequent acquisition of a similar set of statistics for the following year – that is, 1999 – we could use these additional data to perform an 'out-of-sample' prediction test. (We preferred an out-of-sample prediction test over a – perhaps more commonly used – within-sample prediction test, as the latter tends to overstate the predictive ability of any regression model and should thus be avoided, if possible. Also, we did not include Coastguard statistics for the years 2000 and beyond in our prediction test, as that is the period when a number of the coordination centres were closed – giving rise to at least the possibility of a structural break in our regression model.)

As explained in more detail in the MCA-LRA paper itself, we compared the prediction performance of our LRA model against a simple alternative model, which was based on the 'naïve' assumption that the average failure (or success) rate for each coordination centre would remain unchanged from year to year. Making due allowance for the fact that this prediction test was not necessarily very precise because of the relatively small numbers of fatalities (represented by the variable 'LO') being predicted, we found that the prediction performance of the LRA model was not unambiguously superior to that of the 'naïve' model. Should we, therefore, have applied the parsimony principle and rejected the former – more complex – model in favour of the latter – simpler – model?

The reason why we did not take that particular approach is straightforward. As we have already explained in several chapters of this thesis (including, in particular, Section 8.1 above), the MCA models were intended to help us to uncover significant differences in performance and also, where possible, to give us a better understanding of the cause-and-effect relationships driving these performance differentials. In particular, the LRA model provided insights into the effects on our measure of service outcome of key explanatory variables such as average staff workload – whereas the 'naïve' model provided no such insights. In consequence, the 'naïve' model did not aid our understanding of a central issue in the MCA debate (namely, the trade-off between labour productivity and service outcomes) and was, in that respect, distinctly inferior to the LRA model. In other words, statistical performance was only one of the factors that we took into account when selecting our models; explanatory power was just as important (which fits in well with the Critical Realist position that we outlined in Chapter 7 above).

8.2.1.2.3 Developing a random effects model

As we have already noted, the fixed effects version of the LRA model suffered from the problem of multicollinearity. Therefore, we decided to extend our analysis by developing an alternative form of multiple regression analysis involving the assumption of random effects. This second set of OR models was based on Stochastic Frontier Analysis, which distinguishes between different cross-sectional units through differences in (composite) random disturbance terms rather than fixed (i.e. non-random) dummy variables. This might lead to the question: which one is 'better' – the LRA (fixed effects) model or the SFA (random effects) model? But that question does not necessarily have a straightforward answer as there are advantages and disadvantages on either side. For instance, the SFA model has an advantage over the LRA model in that it is less affected by the multicollinearity problem; while it also has a disadvantage in that it does not take full account of the binary nature of our preferred measure of service outcome. Therefore, rather than selecting one kind of model over the other, we examined them both and checked whether their results were broadly comparable. As we have demonstrated in the MCA-LRA paper, that turned out to be indeed the case – increasing our confidence in our overall modelling approach.

8.2.1.2.4 Subjectivity in statistical modelling

To conclude these reflections on the MCA-LRA paper, we shall discuss a more general concern; namely, the extent to which the modelling approach followed in a particular study can ever be wholly independent from the subjective choices made by the modellers concerned. Magnus and Morgan (1999) conducted a couple of practical experiments to assess different ways of developing econometric models. In one experiment, an apprentice modeller was asked to analyse the same data set three times, each time following the published guidance of a different expert modeller. In all cases, the results were different from each other, and also different from the results produced by each of the expert modellers. In Morgan and Magnus's own interpretation, this demonstrates the importance of tacit knowledge in econometric modelling.

In another experiment, Magnus and Morgan asked eight expert teams of modellers, from different universities, to produce statistical estimates of the income elasticity of the demand for food; these estimates were to be produced on the basis of the same data sets, but with each team using its own particular modelling approach. The elasticity estimates produced by the different teams ranged from around 0.38 to around 0.74; as Mingers (2006a) notes, these estimates are clearly vastly different, especially when remembering that they were produced on the basis of the same sets of data! In a more general comment on the lessons from the Magnus and Morgan experiments, Mingers (*op. cit.*, p. 209) argues that their applicability extends to the wider field of statistical modelling:

"... it is clearly the case that experienced modellers could easily come up with significantly different models based on the same set of data, thus undermining claims to researcher-independent objectivity."

The above argument from Mingers is highly relevant to our own approach to statistical modelling, as described in the MCA-LRA paper (and, where applicable, in the other three papers under consideration). As explained earlier in this section, given our desire to gain a better understanding of the cause-and-effect relationships underlying the performance of the Coastguard service, we considered the explanatory power of a model to be as important as its statistical fit. Hence, we preferred our LRA model to the naïve model based on fixed average failure rates, even though the prediction performance of the former model was not unambiguously superior to that of the latter model. But that is essentially a subjective choice on our part, which can only be convincingly justified by our particular emphasis on opening up the black box of the transformation process.

On the other hand, when introducing the Length of Coastline as an explanatory variable into our LRA model, there was no theoretical reason (in terms of the model's explanatory power) to use the non-logarithmic form ('LC') rather than the logarithmic form ('lnLC'). Nevertheless, we preferred the non-logarithmic form because it resulted in a superior statistical fit. But, again, that constitutes a subjective choice, as other researchers might have included the Length of Coastline variable in its logarithmic form – in order, perhaps, to treat it in the same way as the other explanatory variables ('OIR' and 'PAR', which are both included in their logarithmic form). All this is not to say, of course, that 'anything goes'. On the contrary, researchers – including OR modellers – should clearly state on what considerations model-selection decisions are to be based, whether it be statistical fit, explanatory power or any other criteria directly related to the purposes and intentions behind the particular modelling effort. But the Magnus and Morgan experiments (in which statistical fit was a main criterion) suggest that any notion of 'researcher-independent objectivity' is unlikely to be fully achievable in practice.

8.2.2 Final reflections on the MCA-DEA paper

Next, we shall consider the MCA-DEA paper; in particular, the use of the Data Envelopment Analysis technique. The opening sections of the MCA-DEA paper contain a brief explanation of the main ideas behind DEA and also examine in some detail both its advantages and disadvantages compared to the application of regression models, such as SFA. We shall return to the question of how effective DEA turned out to be in the MCA case; but we shall first outline our overall rationale for calculating technical efficiency scores for individual Coastguard coordination centres over time.

8.2.2.1 Overall rationale for calculating technical efficiency scores

A central issue in the MCA case was the rationale for the Government's policy of closing a number of Coastguard coordination centres; and, more particularly, on what grounds particular centres had been selected for closure and others had not. As already explained in Section 5.5 above, in his 1999 review of the Government's closure decisions, Lord Donaldson had identified a key trade-off between the level of staff experience (which may be increased by the closure of a centre as the workload is divided between the remaining ones, with a positive effect on the proportion of lives saved) and their level of local knowledge (which may be reduced by the closure, with a negative effect on the proportion of lives saved). Roughly speaking, that would suggest that closure decisions should focus on those coordination centres where the average staff workload (and the resulting opportunity for developing and maintaining staff experience) was relatively low and the length of coastline covered (as a possible proxy for the amount of local knowledge required) was not already relatively long.

But Lord Donaldson's line of argument did not take account of two important points:

(1) It needs to be statistically established that variables such as the average staff workload and the length of coastline covered actually matter to the Search

And Rescue performance of individual coordination centres. Although Lord Donaldson's report in 1999 had made no real attempt to do so, we succeeded in the MCA-LRA paper in demonstrating the statistical significance not just of these two variables but also of a key environmental variable ('PAR', indicating the average scale of SAR incidents) not even considered by Lord Donaldson.

(2) The possibility cannot be excluded on a priori grounds that coordination centres with relatively low staff workloads and/or long coastlines and/or small-scale incidents might perform better than expected because of other factors (e.g. managerial expertise or staff motivation) that were not explicitly taken into account either by Lord Donaldson or in our own LRA model.

One of the main reasons for constructing the SFA model was to investigate as thoroughly as possible the second point above. (Our wish to develop a random effects alternative to the fixed effects LRA model was another important reason.) As discussed in the conclusions to MCA-LRA paper, even by the Government's own criteria of low staff workloads, it was not entirely clear why particular coordination centres had been selected for closure. From the three centres to be closed, Pentland had the lowest staff workload of all 21 centres, combined with an average length of coastline. But while Oban had a lower-than-average staff workload, it already covered the longest coastline of all 21 centres. And while Tyne Tees had one of the shortest coastlines, its staff workload was only just below the average for all 21 centres. Therefore, whereas Pentland was indeed an obvious candidate for closure, the same could not be said so readily for Oban and Tyne Tees. Moreover, the technical efficiency scores from the SFA model showed that only one of the three soon-to-be-closed centres – namely, Pentland – appeared to perform worse than expected given its staff workload, length of coastline and average scale of incidents, with Tyne Tees being an average performer and Oban actually being one of the better performers in that respect.

The MCA-DEA paper was intended, *inter alia*, to provide a means of comparison between the technical efficiency scores resulting from the application of the SFA technique (which constitutes a parametric method of Frontier Analysis) and the DEA technique (which is a non-parametric method). However, rather than constructing a single DEA model, two different models were examined: one DEA model incorporating the length of coastline ('LC') variable as a proxy for local knowledge (i.e. similarly to our SFA model) and the other DEA model excluding the length of coastline as a relevant variable (which may be more in line with the Government's – although not necessarily Lord Donaldson's – seemingly 'aspatial' perspective of Coastguard SAR performance). But unlike the SFA model, which provided a unique efficiency score for each coordination centre in each time period, the DEA models tended to result in a surfeit of coordination centres with 100% efficiency scores, making comparisons more difficult than would otherwise be the case. However, this is a long-standing problem in DEA modelling, which has been well recognised in the academic literature. We shall, therefore, proceed by reviewing a number of papers that suggest various ways for dealing with this problem.

8.2.2.2 Brief review of relevant papers written by other authors

As we have noted in the MCA-DEA paper, DEA calculates the technical efficiency of each decision making unit (DMU) as the ratio of the weighted sum of outputs to the weighted sum of inputs. In the original version of DEA (as applied in the MCA-DEA paper), all of these weights are allowed to vary freely between zero and one. But because DEA is designed to find the most favourable set of weights for each DMU individually, one may, in practice, end up with a large number of fully (i.e. 100%) efficient DMU's – which is seen by many DEA practitioners as a problem of lack of discrimination. Suggestions for resolving this problem began to appear relatively early in the history of DEA (which was only introduced by Charnes *et al.* in 1978). Notable examples of such early contributions are the papers by Dyson and Thanassoulis (1988), Charnes *et al.* (1990), Wong and Beasley (1990) and Cook *et al.* (1992).

Dyson and Thanassoulis (*op. cit.*) developed a method for setting lower bounds on the DEA output weights based on the consensus view of the stakeholders involved, but only for single-input multi-output or single output multi-input cases. Charnes *et al.* (1990, *op. cit.*) suggested a different approach, whereby transformed (i.e. artificial) input-output data are used to simulate restrictions on the DEA weights. In the particular case considered by Charnes *et al.*, this involved selecting some favoured DMU's as models for the others. Wong and Beasley (*op. cit.*), on the other hand, explored the application of weight restrictions direct to the original data. But rather than restricting the actual DEA weights, the proportion of the weighted sum of outputs of any given DMU devoted to a particular type of output – i.e. the 'importance' attached to that particular type of output by the given DMU – is restricted to a limited range on the basis of expert opinion. (A similar restriction could be set on the importance of a particular type of input for any given DMU, again based on expert opinion.) Cook *et al.* (*op. cit.*) were among the first to use weight restrictions in order to evaluate the relative strengths of different DMU's all lying on the efficiency frontier. Here the main question is how 100% efficient DMU's can be ranked on the basis of their ability to assign a balanced set of weights. Allen *et al.* (1997) provide a useful review of the academic debate (including the papers discussed above) on the use of weight restrictions in DEA and make a number of important points. In particular, they conclude that weight restrictions are meant to encapsulate value judgements in the assessment of relative performance, but there is no all-purpose method for accomplishing this.

Even when more recent contributions to the debate are taken into account, the conclusion drawn by Allen *et al.* still stands in its essence. For example, Podinovski (2005; 2007) states that there is still no consensus in the DEA literature as to what weight bounds actually mean. He claims that the variety of techniques for incorporating value judgements into DEA through weight bounds, each of which is defendable in its own way, leads to confusion and undermines the confidence in the resulting efficiency ratings. Podinovsky's own suggestion is to transform the DEA model to a special form in which the terms introduced by the weight bounds can be interpreted as the trade-offs between the inputs and outputs. (But he also acknowledges that some of his own guidelines may prove to be difficult, if not impossible, to apply in certain practical studies!) Yet another way for improving the discrimination between DMU's in DEA is proposed by Gonzalez-Bravo (2007), whose method involves the construction of several simplified DEA models based on a 'prior-ratio analysis' (PRA). The PRA is meant to start with an initial ranking of the DMU's according to some single-input single-output ratios. The resulting information is then used to construct simplified DEA models that combine different input-output sets as variables. The parallel analysis of these simplified models should then highlight both similarities and differences between the rankings of any given DMU under different model specifications, and identify aspects that would otherwise have gone unnoticed. Again, it would appear that this method depends quite heavily on the (subjective) judgements of the DEA practitioners involved (and perhaps the other stakeholders as well).

A recent overview of key issues in the above debate is provided by Podinovski and Thanassoulis (2007). They start by noting that the lack of discrimination between DMU's – when most of them attain the maximum or near-maximum efficiency score – is still one of the most common problems faced by DEA practitioners. They then proceed to explore a number of options to improve discrimination on efficiency in DEA; including: (1) increasing the number of DMU's; (2) reducing the number of inputs and outputs; (3) using weight restrictions (with the identification of production trade-offs as one possible way for specifying weight restrictions); (4) introducing 'unobserved' DMU's; and (5) developing 'hybrid returns-to-scale' models based on the assumption of 'selective proportionality'. There appears to be no simple way for choosing any option in preference to the others – if any of these options should be selected at all! In general, it may be best to follow Podinovski and Thanassoulis' general advice to check whether the DEA model correctly reflects one's overall understanding of the underlying transformation process, and to incorporate only information on preferences over inputs and outputs that one is fully sure about.

8.2.2.3 Aspects of OR modelling

In the theoretical part of the MCA-DEA paper we posed a number of key questions, which can be summarised as follows. On what information should the specification of the list of inputs and outputs of a DEA model be based? In particular, which environmental variables are significant enough to be included? And should these environmental variables be included as non-discretionary inputs or outputs? In the rest of the MCA-DEA paper we demonstrated the practical application of the particular modelling approach that we had developed to answer questions of this sort. Briefly, our modelling approach was based on the application of appropriate statistical techniques in order to inform the formulation and subsequent analysis of a limited number of DEA models In the MCA case, we first developed multiple regression models in the form of our LRA and SFA models; and then, on the basis of the understanding thus gained, we formulated two DEA models – one including the length of coastline as a relevant non-discretionary output and the other excluding it. Through the parallel analysis of the two DEA models, we wanted to highlight both similarities and differences between the rankings of any given coordination centre under different model specifications – which is not that different in essence from the general approach subsequently recommended by Gonzalez-Bravo (2007).

8.2.2.3.1 Using ratio-based variables

In our regression analysis, the dependent variable ('ODDS') and two out of the three explanatory variables (namely, 'OIR' and 'PAR') took the form of ratios of other variables. For instance, the average staff workload ('OIR') is calculated as the ratio of the annual number of incidents ('IN') divided by the average number of staff employed ('WA'). As the formulation of our DEA models was directly linked to the outcome of the regression analysis, the same was true for the equivalent list of DEA inputs and outputs. As shown in Figure 1 (at the bottom of page 894) of the MCA-DEA paper, the first DEA model included 'OIR' as a discretionary input and 'PAR' as a non-discretionary input, with 'ODDS' as a discretionary output. But, as already explained, all three of these DEA variables had originally been constructed as ratios. For the second DEA model, the variable 'LC' (which, as the length of coastline, is not a ratio) was added as a non-discretionary output.

As we have readily acknowledged in the MCA-DEA paper itself, the use of a ratio-based variable such as 'OIR' (the annual number of incidents per member of staff in a given year; i.e. the average annual staff workload) rather than a more straightforward volume measure such as 'IN' (annual number of incidents; i.e. the total annual workload) is uncommon in DEA studies. In support of this latter assertion we could refer to well-known textbooks on DEA, such as Cooper *et al.* (2006), or to more general textbooks on efficiency analysis, such as Coelli *et al.* (2005), neither of which include ratio-based variables in their DEA examples. But using 'IN' (the total workload) rather than 'OIR' (the average staff workload) in our particular DEA models would have been problematic for the following reasons. As we have pointed out in the relevant section of the MCA-DEA paper, using 'IN' as a simple volume measure for the total SAR workload of a coordination centre would ignore the important fact that in some centres far more staff are employed than other centres. Of course, we could try to take that into account by using 'IN' as a (non-discretionary) output and, simultaneously, 'WA' (the average number of staff employed) as a (discretionary) input. But then, according to the rules of DEA, coordination centres could become efficient merely by achieving a high level of labour productivity, even if none of their efforts resulted in lives being saved – which would go against the whole logic of our argument (namely, that coordination centres should be primarily distinguished by their proficiency in saving lives). Therefore, we maintained that our particular DEA models should not include simple volume measures (such as 'IN'), but 'normalised' volume measures in the form of ratio-based variables (such as 'OIR'), instead.

Note that in the MCA-DEA paper, we advanced two supplementary arguments against 'deconstructing' a ratio-based variable such as 'OIR' into its constituent components 'IN' and 'WA'. Both of these arguments centred on our desire to gain a better understanding of the cause-and-effect relationships driving the performance differentials between coordination centres (in other words, the transformation process underlying our DEA models). First, the average staff workload ('OIR') had already been reported by the MCA as an appropriate measure for the 'busyness' of a coordination centre; and, more importantly, it also played a key role in Lord

Donaldson's trade-off between staff experience and local knowledge – the precise terms of which were, as far as we were concerned, a key area for investigation. Second, returning to our overall approach for formulating the DEA models, the 'OIR' variable first appeared in our LRA model, which – as explained in Section 8.2.1.2 above – we preferred to alternative statistical models as much for its explanatory power as its statistical performance. In short, both of these arguments still stand.

8.2.2.3.2 Time inconsistency in DEA

Having dealt with the question of how we formulated our DEA models, we can now review the results of the DEA analysis. As discussed in the MCA-DEA paper, we computed the results for each of the two DEA models for each of the four years under consideration. We found that the 'within-year' technical efficiency rankings of individual coordination centres were often subject to substantial fluctuations from year to year. For example, in DEA Model 1 (without 'LC' as a non-discretionary output), the within-year technical efficiency of the Aberdeen centre fluctuated as follows: 1995 – efficiency 94.8%, ranked 6th; 1996 – efficiency 100.0%, ranked joint 1st; 1997 – efficiency 67.0%, ranked joint 11th; 1998 – efficiency 60.3%, ranked 16th. And for DEA Model 2 (with 'LC' as a non-discretionary output), the picture for Aberdeen was very similar: 1995 – efficiency 95.9%, ranked 8th; 1996 – efficiency 100.0%, ranked joint 1st; 1997 – efficiency 70.2%, ranked 15th; 1998 – efficiency 60.4%, ranked 16th. This problem of 'time inconsistency' is not unusual when DEA is applied to panel data rather than cross-sectional data for a single time period; an interesting practical example is provided by Hannesson (2005). In the MCA case, this problem is not helped by the fact that the annual numbers of lives lost per coordination centre tend to be relatively low (at least compared to the number of persons rescued). We also examined the 'between-year' technical efficiencies of the coordination centres for indications of a consistent time trend. But, in this respect at least, the DEA results matched the results from the earlier SFA model. That is, we found little evidence for a consistent shift in technical efficiency over the years, although the average between-year efficiency in the last year (1998) was in both

DEA models higher than in the first year (1995) – which could perhaps be taken as some sign of general improvement.

8.2.2.3.3 The question of weight restrictions

As expected – since we applied DEA in its original, unconstrained form – we found in each year multiple coordination centres with 100% within-year technical efficiency. In DEA Model 1 (without 'LC'), the numbers of 100% efficient centres were as follows: 1995 – 4; 1996 – 8; 1997 – 4; 1998 – 8. And in DEA Model 2 (with 'LC'), the situation was – again as expected (since DEA Model 2 contains an additional output variable) – even more pronounced: 1995 – 6; 1996 – 9; 1997 – 7; 1998 – 11. Nevertheless, we decided against using either weight restrictions or some other method suggested by the relevant academic literature (e.g. Podinovski & Thanassoulis, 2007) to resolve this lack of discrimination. Our reluctance to introduce weigh restrictions into the DEA models was inspired by two considerations. First and foremost, we simply did not have the information that would be required to make the value judgements which the weight restrictions are meant to encapsulate. Trying to make such value judgements without the necessary knowledge about the preferences of the main stakeholder groups (including not just the Government and the management of the MCA, but also MCA staff and their trade union, local volunteers, various groups of service users, etc.) would have introduced an unacceptable degree of subjectivity on our part into the analysis. Second, we felt no real need to introduce weight restrictions into the DEA models, since we had already performed a form of Frontier Analysis in which strict weight bounds are naturally imposed (through statistical estimation); namely, SFA.

8.2.2.3.4 Comparing SFA and DEA results

One of our intentions for the MCA-DEA paper was, as explained in Section 8.2.2.1 above, to provide a means of comparison between the SFA and DEA efficiency scores. In the Conclusions (on page 640) of the MCA-LRA paper, we had already stated that "it is well-known from the relevant academic literature that the results derived from the application of different theoretical approaches to performance measurement (such as SFA and DEA) are not necessarily consistent with each other". While we provided a couple of references to support our claim, one very useful reference was not yet available when we completed the MCA-LRA paper; namely, Chapter 7 (entitled "A comparison of SFA and DEA") in the textbook by Jacobs *et al.* (2006). The latter authors give a list of reasons why SFA and DEA produce different efficiency estimates – too many to discuss in detail here, although some of the most important ones are analysed in the theoretical part of the MCA-DEA paper. They then illustrate the severity of these differences with a practical example involving the analysis of a cross-sectional data set for acute hospitals in the English NHS. In their conclusions to their Chapter 7, Jacobs *et al.* (*op. cit.*, p. 165) summarise their view on the SFA-versus-DEA debate as follows:

> *Ultimately ... there is no consensus in the literature on the 'best method' for estimating the efficiency frontier. Some commentators have argued that consensus is not necessary, as long as a set of consistency conditions is met (Bauer et al., 1998). To the extent that there is no a priori reason to prefer one technique over the other, it seems prudent to analyse efficiency using a broad variety of methods to 'cross-check' the results (Stone 2002). Bauer et al. (op. cit.) argue that the efficiency estimates should be consistent in their efficiency levels (with comparable means, standard deviations and other distributional properties), consistent in their rankings, consistent in their identification of best and worst performers, and consistent over time. Rarely, however, are these consistency conditions likely to be met, as is the case for the data analysed in this chapter."*

In the MCA-DEA paper, in order as much as possible to compare like with like, we compared the SFA technical efficiency scores to those from DEA Model 2 (that is, the model including the Length of Coastline variable). To mitigate the problem of time inconsistency (as explained in Section 8.2.2.3.2 above), we computed the SFA and DEA efficiency scores on the basis of the average values of all relevant variables over the four-year period 1995-98. When comparing the rankings produced by SFA and DEA, respectively, we found – not surprisingly, in view of the above quote from Jacobs *et al.* – that the value of Spearman's ρ rank was rather low ($\rho = 0.355$, with a p-value of 0.115). However, as we have demonstrated in the relevant section of the MCA-DEA paper, the difference in the efficiency rankings was strongly magnified by the 100% efficiency scores in the DEA model for the Dover, Pentland, Stornoway and Shetland coordination centres, respectively, all of which were due to extreme values in one or other of the inputs and/or the non-discretionary output. If these four centres were to be excluded from the rankings, then the value of the rank correlation coefficient would become highly significant ($\rho = 0.827$, with a p-value of 0.000). From this analysis, it would appear that our own SFA and DEA efficiency estimates were not quite as lacking in consistency as might have been predicted from the remarks by Jacobs *et al.* (*op. cit.*). But any success in that respect was not least due to our decision to use four-year averages rather than annual data in calculating the SFA and DEA efficiency scores used for comparison.

8.2.2.3.5 The overall role of the SFA and DEA models

To conclude these reflections on the MCA-DEA paper, we shall briefly review the role of the two Frontier Analysis techniques, SFA and DEA, in the wider context of the MCA studies. As we have already explained on a number of occasions, the MCA models were generally intended to help us to uncover significant differences in performance and also, where possible, to give us a better understanding of the cause-and-effect relationships driving these performance differentials. More particularly, as we have noted in Section 8.2.2.1 above, even by the Government's own criteria of low staff workloads, it was not entirely clear why particular Coastguard coordination centres had been selected for closure. Therefore, we wanted to use Frontier Analysis techniques to investigate whether coordination centres with relatively low staff workloads and/or long coastlines and/or small-scale incidents might perform better (or worse) than expected because of other factors (e.g. managerial expertise or staff motivation) that had not been explicitly taken into account in earlier work. In the end, we found that the DEA Model 2 results broadly confirmed the earlier SFA results for the Oban centre (better-than-average performer) and also the Tyne Tees centre (average performer) – although not for the Pentland centre, which changed from being one of the worst performers under SFA to being one of the best under DEA because of its extremely low value for the 'OIR' variable. (Note that there is an error in the MCA-DEA paper on this particular point! At the top of page 899 in the MCA-DEA paper, it states that Pentland had the lowest average value for the 'PAR' variable – but this should have read the 'OIR' variable, instead.)

Whatever the merits of the overall logic underlying the Government's closure plan (as articulated, in particular, by Lord Donaldson), it would appear that, apart from the decision to close the Pentland coordination centre, our Frontier Analysis results did not tend to support the particular selection that the Government had made. However, we would agree – not just in view of the obvious conflict between the SFA and DEA efficiency scores for the Pentland centre, but also more generally – with the important advice given by Jacobs *et al.* (2006, p. 165), that:

> *"... [T]he efficiency scores derived from SFA ort DEA should not be interpreted as accurate point estimates of efficiency, and it would be inappropriate to take action solely on the basis of these estimates ... Rather, where estimates of relative efficiency are obtained, these might be used as signals about where to direct more investigative energy."*

Although we would still contend that our performance measurement framework is not just more complex but also more realistic than the one apparently used by the Government – with 'realistic' to be understood in the sense of having explanatory power through its ability to conceptualise fundamental mechanisms (such as the effects of staff experience and local knowledge) in the real domain – the results from our SFA and DEA models were not meant to be taken as definitive evidence against, or in favour of, certain policy decisions. Instead, we wanted to employ these Frontier Analysis techniques primarily in an exploratory fashion. However, the development of our SFA model was probably a more important step in that regard, as the DEA models did not yield many additional insights beyond confirming some, but not all, of the earlier SFA results. (And where the DEA efficiency scores were inconsistent with the SFA ones, as in the case of the Pentland coordination centre, we had reason to consider the SFA efficiency scores to be more informative, as they were less affected by extreme values for the input variables.)

8.2.3 Final reflections on the MCA-BBN paper

Unlike the extended nature of the discussion on the MCA-LRA and MCA-DEA papers in Sections 8.2.1 and 8.2.2, respectively, we shall limit our final reflections on the MCA-BBN paper to a review of the rationale for developing our BBN model and its contribution to a greater understanding of the transformation process in the MCA case; and we shall also briefly consider the role that BBN's could play more generally in OR modelling for performance measurement.

8.2.3.1 The rationale for developing the BBN

As already explained in Section 6.2 above, in our LRA model we had used aggregate Coastguard statistics to estimate the significance of 'macro' factors – namely, the average scale of incidents, the average workload of watchkeeping staff and the length of coast line monitored – in explaining the annual average success rate (in terms of the proportion of lives saved) for each Coastguard coordination centre. Based on these macro factors, we then developed SFA and DEA models to analyse the performance differentials between individual coordination centres. But as Jacobs *et al.* (2006) noted, one of the main outcomes of Frontier Analysis should be to indicate the particular areas to which more investigative energy ought to be directed. The importance of the average staff workload had already been recognised by Lord Donaldson (1999). But why did the fact that coordination centres like Stornoway tended to cope with relatively large-scale incidents constitute a positive factor in their local operating environment; and why was the opposite true for centres like Dover who usually had to deal with much smaller incidents? And was the long length of coastline covered by the Oban centre really a good indicator for the amount of local knowledge required by its watchkeeping staff?

The answers to questions such as these were unlikely to be found in the aggregate Coastguard statistics. Instead, we realised that we would have to analyse the detailed records of individual incidents – assuming that we could get access to them. However, even if we could get access to at least some of these records, we still had to learn how to analyse them properly. We were already aware that a large of number of details was recorded for each incident – and not always accurately, as explained in Section 2.4 above. Many of these data might have been more or less irrelevant to the chances of success for the SAR operation concerned; but some other data could actually represent important micro factors at work. In short, we needed to find a suitable way to determine which detailed risk factors were likely to be important at the level of individual SAR incidents, how each of these risk factors might interact with the others and, finally, how such risk factors tend to manifest themselves in the form of explanatory variables (such as 'OIR' and 'PAR') at the macro level as addressed by the LRA, SFA and DEA models. (As we have noted in Section 6.2, we were not so much trying to justify the selection of explanatory variables for the macro models that we had developed earlier, as wanting to gain a better understanding of the meaning of these aggregate variables and the possible reasons for their estimated effects on the average success rate for each Coastguard coordination centre.)

When, eventually, we gained access to a set of individual incident records at MCA headquarters and were also given the opportunity to interview some senior MCA managers, we used the resulting information to start developing a micro-level model of the primary variables contributing to an effective SAR operation in the form of a Bayesian Belief Network. We preferred the methodology of BBN's over, say, the construction of a more loosely structured cognitive map and/or influence diagram for a number of reasons. First of all, given their professional standing and experience, the two MCA managers to be interviewed could fairly be regarded as experts in the field of maritime SAR operations. Therefore, we wanted to follow a well-established and rigorous process that was designed to minimise as much as possible the effects of various biases that are known to affect expert judgements in such situations. (The process followed in constructing a cognitive map or influence diagram from interview data might not be as rigorous in practice.) Second, our statistical analysis of SAR incident data – that is, the aggregate Coastguard statistics and subsequently also some of the individual incident records – enabled us to exercise some level of empirical control over the quantification of our BBN. (We put the results from our statistical analysis in the form of contingency tables, which, by their nature, could be applied much more readily to a BBN than to a more loosely structured influence diagram.) Third, the methodology of BBN's had already been successfully applied to risk-analysis studies in some other areas (Sigurdsson *et al.*, 2001) and an extension to maritime risk seemed quite natural. (Since we have already provided a short review of papers on maritime risk in Section 8.2.1.1, no further discussion of the literature on that topic is necessary here.)

8.2.3.2 The contribution of the BBN to greater understanding

The main contribution of the BBN was to show how each of the three macro factors identified in our LRA, SFA and DEA models – namely, the average scale of incidents, the average workload of watchkeeping staff and the length of coast line monitored – could represent a somewhat crudely-measured 'proxy' for a set of underlying micro factors. (For the rest of this paragraph and the next, please refer to Figure 4 (on page 945) and Table 2 (on page 946) of the MCA-BBN paper.) First, the average scale of incidents may be an indication of the likely severity of an incident. The latter was judged by the MCA experts to be one of the high-impact variables in the BBN. Second, the average workload of watchkeeping staff may be an indication of factors such as the competency of the watchkeeping staff (as higher workloads may contribute to greater experience of watchkeeping staff through a learning-curve effect), the likelihood of other incidents taking place concurrently, and perhaps also the presence of effective communications. Third, the length of coastline monitored by a coordination centre may be an indication of more localised factors; in particular, whether the location of the casualty is known. But knowledge of the location, which was judged by the MCA experts to be another one of the high-impact variables in the BBN, is directly affected by local knowledge; and – as we have already noted on numerous occasions in this thesis – the acquisition of local knowledge by watchkeeping staff may be hampered by the length of coastline to be monitored. Another localised factor may be the competency of SAR units (consisting of local volunteers and others), which is influenced by the distance of SAR units from the incident – and the length of coastline monitored may tend to have an adverse effect on that distance.

Note that not all of the parent variables of the leaf node "Effective SAR Operation" in the BBN could be linked to the macro factors that we had identified earlier. In particular, the important roles of the variables Casualty Situation, which was judged by the experts to be another high-impact variable, and Favourable Weather may be only dimly reflected in the explanatory variables that were identified at the aggregate level. (The effect of the Favourable Weather variable might well show up more clearly at the aggregate level, if the UK Coastguard Statistics were routinely published on a quarterly, rather just than an annual, basis. But that is not the case, unfortunately). More generally, as we have already acknowledged in Section 7.4, our efforts at triangulating our empirical findings from the macro and micro levels of analysis are not yet complete. Because the MCA experts struggled to provide quantitative estimates of the impacts of the various micro factors, we still intend to collect and analyse statistically a larger set of individual incident records in order to strengthen the BBN's empirical basis and explicate further its links with the macro level of analysis. (As before, the BBN lends itself much more readily than a more loosely structured influence diagram to the application of the contingency tables resulting from the statistical analysis of additional incident records.)

8.2.3.3 The wider role of BBN's in OR modelling for performance measurement

In terms of the different categories of OR modelling approaches that we developed in Section 5.3 above, the BBN represents a 'category 3' model. This category consists of models (usually, but not always, stochastic) that involve the application of some kind of simulation, or other non-optimising analytical, method. Although such models would not appear to be directly concerned with performance measurement, unlike 'category 4' models such as SFA and DEA, they are strongly related to the latter category of models in that they can be included under the common heading of "models for evaluating effects of changes in systems" (Sanderson & Gruen, 2006). Whereas SFA and DEA models help us to uncover significant differences in performance, a BBN could give us a better understanding of the cause-and-effect relationships driving these performance differentials.

In the context of the MCA studies, the role of the BBN was to help us understand how differences in aggregate performance between Coastguard coordination centres over time might represent the cumulative effects of risk factors that tend to be at work in individual incidents. In other words, the BBN was meant to provide the micro-foundation for the empirical regularities in SAR performance at the macro level. But it is not always necessary to start off at the macro level of analysis, nor should a BBN only appear in the later stages of a performance measurement project. In the context of other OR studies, particularly those with a primary focus on risk analysis at the micro level, constructing a BBN could be one of the first steps in the modelling process. This could then be followed by the development of models (such as SFA and DEA) that allow the OR practitioner to contrast how well different decision making units manage the risks analysed in the BBN.

In general, the potential role of a BBN in OR modelling for performance measurement – whether in the public or the private sector – is very similar to that of a Discrete Event Simulation (DES) model. Both are essentially micro-level 'category 3' models, which are used to model the transformation process underlying the organisational performance measured more directly by the 'category 4' models, with the distinction that a BBN is often developed in the specific context of risk analysis and a DES model can be used for more general process analysis. In Section 8.2.4 below, we shall briefly review the contribution of DES models to performance measurement in a public service organisation such as the NHS and many of our remarks on that issue also apply to the use of BBN's.

8.2.4 Final reflections on the MSU-DES paper

As part of our final reflections on the MSU-DES paper, we shall first briefly review a number of recent papers written by other authors on aspects of simulation modelling that are relevant to issues considered in our paper. This will be followed by some reflections on the contribution that DES models can make to performance measurement in public service organisations.

8.2.4.1 Brief review of some recent papers on simulation modelling written by other authors

The MSU-DES paper already contains a number of references to articles that describe the use of DES models to investigate the two key aspects of managing a health care system; namely, the management of patient flows and the allocation of the required resources. The academic literature on this topic is vast and still expanding. In this section, we shall limit ourselves to a review of a representative selection of articles that appeared in the Journal of the Operational Research Society after our own MSU-DES paper had been published there. In particular, we shall quickly cover a couple of papers by Robinson (2008a and 2008b) on conceptual modelling for simulation; a series of papers by Cooper *et al.* (2007), Proudlove *et al.* (2007), Sachdeva *et al.* (2007) and Eldabi *et al.* (2007), respectively, on simulation modelling of health care systems; and a paper by Taylor and Dangerfield (2005) on the modelling of feedback effects in health care systems.

8.2.4.1.1 Conceptual modelling for simulation

In Section 4.2 above, we have already quoted Pidd's (2003) generic definition of an (OR) model; and we highlighted the importance of distinguishing between different motivations for OR modelling, referring to Mitchell's (1993) distinction between the idea of a model as 'a statement of the beliefs held relevant to the issue under study' and the somewhat different idea of a model as 'a device, usually for making predictions'. In the specific context of simulation modelling, Robinson (2008a, p. 283) provides his own definition of a conceptual model, as follows:

> *"The conceptual model is a non-software specific description of the computer simulation model (that will be, is or has been developed), describing the objectives, inputs, outputs, content, assumptions and simplifications of the model."*

Note that, unlike Pidd, Robinson does not mention the possible reasons for constructing a model in his definition; but, as we shall explain in the next paragraph, that does not at all mean that he considers modelling objectives to be unimportant. In the rest of his paper, Robinson describes the requirements of a conceptual model in terms of validity, credibility, utility and feasibility. As we have already provided an extensive discussion of model validation in Chapter 7 above, we shall not pursue that particular issue here – suffice it to say that Robinson's four requirements (or validation criteria) are not necessarily inconsistent with our own views on such matters as expressed in Section 7.3. Robinson also contends that the need to 'keep the model simple' is the overarching requirement in conceptual modelling and refers approvingly to the first of Pidd's (*op. cit.*) general principles of OR modelling ('model simple, think complicated') in that regard. (We have already listed Pidd's general principles in Section 5.4 above and shall reflect in Section 9.1 below on the extent to which we have been able to apply them in the MCA and MSU studies.)

In his follow-on paper, Robinson (2008b) describes his practical framework for conceptual modelling, the steps of which could be considered to be fairly conventional: understanding the problem situation; determining the modelling and general project objectives; identifying the model outputs (responses); identifying the model inputs (experimental factors); and determining the model content (scope and level of detail), identifying any assumptions and simplifications. His conclusion, however, is arguably more interesting; namely, that the notion that there is a right conceptual model for any specified problem is false. He gives two specific reasons for this claim. First, conceptual modelling is an art and different modellers will not come to the same conclusions. Note the clear echoes of the findings by Magnus and Morgan (1999) that we discussed in Section 8.2.1.2.4 above! Second, any model is the result of an agreement (or compromise) between the relevant stakeholders (including the modellers, clients and domain experts), each of which has his or her own preferences for, and perceptions of, what is required. Robinson (2008b, p. 303) summarises his conclusion – with which we can only agree – as follows:

> *"In short, there is no absolutely right conceptual model, because the model is dependent on the preferences and perceptions of the people involved in the simulation study. It would seem that the idea of developing conceptual modelling frameworks that will always lead to a single best model is futile. Instead, our aim should be to provide frameworks that provide a means for communicating, debating and agreeing upon a simulation conceptual model, while also releasing the potential for creativity in the modelling process."*

An important point to add is that any such modelling frameworks should conform to the model validation criteria agreed between the key stakeholders.

8.2.4.1.2 Simulation modelling of health care systems

Many of the 14 papers in the February 2007 special issue of the Journal of the Operational Research Society (JORS) devoted to 'Operational Research in health' discuss (often in the form of case studies) interesting aspects of simulation modelling of health care systems. In this section, we shall select four of these papers for a brief review in other to highlight some specific issues that are relevant to our MSU-DES paper – in particular, the extent to which separate OR models should, as a general principle, be kept as simple as possible and the need for multimethodology.

First, the paper by Cooper *et al.* (2007) compares the use of three different OR modelling approaches for evaluating health care interventions; namely, decision trees, Markov processes and DES models. The authors argue that the choice of modelling technique depends on a range of factors, such as acceptance of the technique in question, model 'error', model appropriateness, dimensionality and ease and speed of model development. But, in general, DES allows the OR practitioner to construct more complex, dynamic and interactive models. Proudlove *et al.* (2007), on the other hand, stress the need to keep any OR model as simple as possible, referring – like Robinson (2008a) – Pidd's (2003) maxim of 'model simple, think complicated'. They contend that complex and descriptively powerful DES models may actually have less practical impact, since such models do not provide the 'transferable tools and insights' that are useful to problem owners or policy makers. Third, Sachdeva *et al.* (2007) suggest that OR modelling efforts are more likely to gain acceptance among stakeholders, and thus to result in sustained organisational change, if they are aimed at combining different – hard and soft – OR modelling approaches. (This fits in well, of course, with our discussion of multimethodology in OR modelling in Chapter 6 above.) Finally, the paper by Eldabi *et al.* (2007) attempts to take a more general overview of possible future developments in the simulation modelling of health care systems. Based on a limited survey of simulation experts, they end up arguing – like Sachdeva *et al.* (*op. cit.*) – in favour of joining up different modelling approaches.

8.2.4.1.3 Modelling feedback effects in health care systems

In various chapters of this thesis, including Sections 3.6 and 5.1 above, we have mentioned the potential importance of feedback effects in public service systems transforming inputs into outputs and outcomes. Such effects could take the form of negative feedback control loops, which are generally taken to be indispensable for keeping service performance on target. But – as noted, *inter alia*, by Boland & Fowler (2000) and Bevan & Hood (2006) – if the performance targets set by management are felt as too challenging by the staff providing the service, then positive feedback loops (based on perverse incentives) could unwittingly be created. In that case, evidence of underperformance might induce staff to behave (usually covertly) in ways that tend to reduce (actual, although not necessarily observed) performance even further.

Although in our own studies we noted only a single instance of a negative feedback loop – in the MSU study, as described in Section 3.6 – and no clear evidence of any positive feedback loops based on perverse incentives, the issue is clearly important from the general perspective of performance measurement in public service organisations. The paper by Taylor and Dangerfield (2005) is a good example of how OR modelling can be used to gain a better understanding of such feedback effects in a health care system. Rather than applying DES, the authors of this paper used an alternative simulation approach in the form of System Dynamics (SD). As the name implies, SD is specifically designed for modelling the dynamic behaviour of (service or other) systems; in particular, the analysis of feedback mechanisms. Compared to DES, SD models tend to be more 'holistic', focusing on the macro (aggregate) rather than on the micro level of analysis. According to Taylor & Dangerfield, the main aim of SD is to provide an understanding of the 'basic trends' rather than to produce precise quantitative predictions. Based on two case studies involving the reconfiguration of cardiac catheterisation services provided by particular NHS hospitals, Taylor & Dangerfield constructed SD models to demonstrate why and how increases in service capacity could stimulate demand from patients for the services concerned. In other words, unlike many DES models in which patient demand is treated as a wholly exogenous variable and, given static demand, capacity increases lead to reductions in patient waiting times, SD models can be used to show, at least in broad terms, that demand is to some extent endogenous and unlikely to remain static in such cases, possibly negating much of the positive impact on waiting times in the longer run.

But although SD constitutes a different approach to simulation modelling than DES, both fall in our third category of OR modelling approaches, as we have already explained in Section 5.3. Therefore, like micro-level DES models (and also BBN's, as explained in Section 8.2.3.3 above), macro-level SD models can give us a better understanding of the cause-and-effect relationships driving any significant performance differentials uncovered by 'category 4' models such as SFA and DEA – whereby SD models are particularly useful when organisational performance is likely to be affected by feedback effects in the longer run.

8.2.4.2 The contribution of DES models to performance measurement in public service organisations

We have already stated repeatedly in this thesis that, in terms of the different categories of OR modelling approaches that we developed in Section 5.3 above, a DES model represents a 'category 3' model. And we have also argued that models in this particular category are essentially complementary to approaches such as SFA and DEA that are more directly concerned with measuring differences in organisational performance. However, in order fully to justify the inclusion of the MSU-DES paper in a thesis entitled 'OR Modelling for Public Sector Performance Measurement', it would seem useful at this point to summarise the main reason for this complementarity.

As we have discussed in Section 4.1 above, there are (at least) eight specific goals which managers of public service organisations may wish to pursue; namely to: evaluate; control; budget; motivate; promote; celebrate; learn; and improve. But to foster improvement could be regarded as the core purpose behind the other seven (Behn, 2003). However, for improvement to be feasible, one needs to know 'what works, for whom, when, in what circumstances and why' (Ferlie *et al.*, 2003; Hodgson *et al.*, 2007). And this goes for managers of public service organisations just as much as for managers of any other kind of organisation – except that public service managers face specific organisational problems and challenges that make it particularly hard for them to measure performance effectively, as explained in Chapter 3 above. The main reason why DES, as an example of our third category of OR modelling approaches, is highly relevant to the overall subject of this thesis is that it allows us to examine the cause-and-effect relationships inside the black box of the process transforming inputs into outputs and outcomes in public service organisations. Or, to put it another way, performance measurement – particularly if done for the primary purpose of organisational improvement – hardly makes sense unless one has a suitable tool (such as DES or similar OR modelling approaches) to model the underlying transformation process and thereby to understand the cause-and-effect relationships driving performance. In short, the boundary between performance measurement and performance management is indistinct; and, for the reasons given above, we have chosen to include under the heading of performance measurement OR modelling approaches such as DES, which other researchers might prefer to regard as properly belonging to performance management. (In our own view, performance management is principally concerned with the incentives provided by the organisation's owners and/or managers to induce other stakeholders (primarily, the employees) to act in effective pursuit of the organisational goals – but that is a topic we have barely touched upon in this thesis.)

Our argument in favour of treating cause-and-effect modelling (in whatever form) as an essential complement to performance measurement proper fully matches recent developments in the specialist academic literature on performance measurement – particularly, in terms of one of its most popular manifestations; namely, the 'Balanced Scorecard' approach. In the first of a series widely cited papers in the Harvard Business Review, Kaplan and Norton (1992) presented the Balanced Scorecard principally as a new approach to performance measurement. However, as they increased their experience with the development and implementation of Balanced Scorecard systems in practice, Kaplan & Norton (2004) put more and more emphasis on the role of 'strategy maps' as an essential element of any performance measurement system based on Balanced Scorecard principles. But strategy maps are simply diagrams describing important cause-and-effect relationships by means of connecting arrows. In other words, even performance measurement 'gurus' such as Kaplan and Norton acknowledge that effective performance measurement requires some tool – however, simple in their case – that enables us to understand what actually causes good (or bad) performance.

9 CONCLUSIONS

9.1 The application of general principles of OR modelling in the four papers

After our final reflections in the previous chapter on the four papers under consideration, we can now present our concluding views on the issues raised in this thesis. In particular, we shall seek to provide summary answers to the list of research questions that we raised at the beginning of this thesis (in Section 1.2). However, one of these questions requires a somewhat longer discussion that can, to some extent, be given separately from our treatment of the remaining questions. The question that we are referring to is number 2(b) in our list; that is, how applicable are more general principles of OR modelling within the overall context of performance measurement in public service organisations? It is convenient to give our initial response to that particular question first. The list of general principles of OR modelling that we shall use as our reference point is the one suggested by Pidd (2003), as described in Section 5.4 above. Our task here is to reflect on the extent to which we have been able to apply these general principles in the MCA and MSU studies.

9.1.1 Model simple, think complicated

The fact that this general principle is the first in Pidd's list appears to indicate its importance; a view shared by OR practitioners such as Robinson (2008) and Proudlove *et al.* (2007), as we have noted in Sections 8.2.4.1.1 and 8.2.4.1.2 above. However, the principle was not quite as easy to apply in our own studies.

Both the Maritime and Coastguard Agency and the Musculo-Skeletal Unit represent (highly) complex examples of transformation systems. While in each case we endeavoured to keep the models as simple as possible, the overriding requirement was to meet the respective purposes that we had set for each model and to respond, where relevant, to the needs of our clients. In the MCA case, we started with an aggregate – and, therefore, relatively simple – model of the transformation process underlying the annual Coastguard statistics. But to gain a better understanding of the various risk factors involved, this macro-level model later had to be complemented by a much more detailed – and rather more complex – micro-level model. In the MSU case, our modelling methodology was explicitly based on the idea of requisite models evolving over time (Phillips, 1984). That is, the form and content of each successive simulation model was allowed to evolve in line with the issues that were of concern to our clients.

In short, based on the lessons learnt from our own studies we would accept the need to keep an OR model as simple as possible as an important requirement – but, unlike Robinson (*op. cit.*) for instance, not necessarily an overarching one. For us, the key question in this regard is as follows: how simple is 'as simple as possible'? And the answer to that question depends on the particular purposes that a study is aimed at. As we have concluded in Section 8.1, at the end of our structured comparison between the four papers under consideration, different purposes may require different types of OR modelling approaches, some more complex than others.

9.1.2 Be parsimonious, start small and add

As we have explained in the previous section, the development of the MCA models did conform to this principle. We started at the macro-level of analysis with a relatively small (Binary) Logistic Regression Analysis (LRA) model and used that as a basis for adding Stochastic Frontier Analysis (SFA) and Data Envelopment Analysis (DEA) modelling approaches. (We have already explained in Section 8.2.1.2.2 why the results of the prediction test on the LRA model did not induce us to reject that model in favour of a (very) simple alternative model. Briefly, the reason for our decision on that point is in accordance with our argument in the previous section – that is, for the purposes of the MCA studies the explanatory power of our models was just as important as their statistical performance.) Subsequently, we further expanded our analysis by adding a Bayesian Belief Net (BBN) in order to understand the detailed risk factors operating at the micro level. In the MSU case, our sequence of Discrete Event Simulation (DES) models also corresponded to this principle, although that had not necessarily been our initial intention. As explained in the previous section and the MSU-DES paper itself, the methodological principle in the MSU study was that of requisite modelling. In this particular case, as the questions asked by our clients became more complex, so became the structure of each new version of our DES model.

But one could imagine another situation where the knowledge and understanding gained from the application of complex models would enable a researcher to distinguish more effectively between necessary and irrelevant details and thus to eliminate the latter. This is indeed the underlying idea for 'general-to-specific' modelling in modern econometrics (Campos *et al.*, 2005), in which an initially general model is subject to progressive simplification as long as it adequately characterises the empirical evidence within the researcher's theoretical framework. Therefore, the universal nature of the 'start small and add' principle is not guaranteed in practice, although it happens to describe quite well the overall development of our models in the MCA and MSU studies (although not necessarily the detailed process of specifying and testing our regression models in the MCA case, which proceeded on a general-to-specific basis.)

9.1.3 Divide and conquer, avoid mega models

This principle depends on the issue under investigation having, in some sense, a modular structure. To be more precise, even though reality may be highly complex (as we would expect it to be if we adopt a Critical Realist position, for instance), the aspects of the real-world problem situation that we are interested in must be, at least to some extent, separable. Different models (and methodologies) can then focus on different facets of the issue. This was indeed the case in the MCA and MSU studies. In the MCA case, as explained in the previous two sections, we were able to construct a series of complementary models, each of which was designed to highlight a different aspect of the situation: namely, the LRA model (to gain an initial knowledge of the explanatory factors influencing SAR performance at the macro level), the SFA and DEA models (to uncover significant differences in performance between Coastguard coordination centres) and the BBN (to achieve a more detailed understanding of the risk factors affecting individual SAR incidents). Similarly, in the MSU case we developed a sequence of DES models, each of which was meant to help us answer different questions from our clients.

But the point missing from this principle is that, within each study, all of the different models should still be brought under a common (meta-)theoretical framework. We have expounded the reasons for this in Chapter 6 and (particularly) Chapter 7. In other words, unless a researcher uses the same type of research method throughout an investigation (as we did in our MSU study) or avoids mixing methods across different scientific paradigms, following a 'divide and conquer' guideline implies that potential problems of multimethodology in OR modelling – in particular, paradigm incommensurability – should be confronted and, where possible, resolved.

9.1.4 Use metaphors, analogies and similarities

This principle is cast in very general terms. In our particular interpretation, OR practitioners should make their models as transparent as possible, so that these models can be more easily understood and accepted by the other stakeholders in the issue under concern.

In the MCA case, the construction of the initial LRA model and, perhaps even more so, the subsequent SFA and DEA models involved the application of relatively sophisticated statistical and mathematical techniques, with which most of the other stakeholders (including the management of the MCA, the watchkeeping staff and their trade union representatives, local SAR volunteers, various groups of service users etc.) were not expected to be very familiar. The management of the MCA seemed to be less inclined to accept the validity of these quantitative models, even though the key results – for instance, that 'macro' factors such as the average staff workload and the length of coastline to be monitored tend to have significant, but opposing, effects on the average SAR performance of each coordination centre – had already been foreshadowed by the earlier, qualitative, analysis done by Lord Donaldson (1999).

In contrast, the subsequent BBN offered a couple of advantages: first, relevant stakeholders (in particular, senior MCA managers) could, as domain experts, be directly involved in its construction; and, second, the interconnections between the various risk factors were shown in a clear graphical form, making the model's internal structure more transparent to non-specialists in OR modelling. Similarly, our clients in the MSU case were very much attracted by the visual interactive nature of modern simulation software, such as Simul8, which helped them to understand the internal structure of the simulation models and allowed them to contribute more effectively to the development of these models. Therefore, we would conclude that we have managed to apply this principle in our own studies – at least, in part and in the particular interpretation that we have given to it.

9.1.5 Do not fall in love with data

Although we would agree with the general sentiments behind this principle, its practical application is not necessarily unproblematic. In the MCA case, the annual Coastguard statistics were, at first, the only data available to us; and the highly aggregate nature of these data necessitated the development of a macro model of the underlying transformation process. We were only able to construct a micro model of individual risk factors when we gained access to a set of individual incident records and were also given the opportunity to interview some senior MCA managers. In other words, the specific nature of the data available in this case determined what type of OR model could be built; in particular, whether it took the form of a micro or a macro model (Mitchell, 1993).

In the MSU case, on the other hand, the decision to use a DES modelling approach came first; and the specific structure of each successive DES model determined what data needed to be collected on patient arrivals, the types of treatment required, the activity times and patient routings involved, and the availability of human and capital resources. However, as we explained in Sections 2.3 and (particularly) 8.1 above, we lacked the data to construct an outcome measure based on actual improvements in the health status of eligible patients and were forced to focus on a range of output measures, such as patient waiting times, instead. For the MCA and MSU studies as a whole, we would conclude that the applicability of Pidd's (1993) notion that 'the model should drive the data, not vice versa' was somewhat limited, as problems with data availability continued to play an important role in practice.

9.1.6 Model building may feel like muddling through

We would argue that this principle has been a very good description of our own modelling efforts in the MCA and MSU studies. (Particularly in the MCA case, where the underlying transformation process simply had not been modelled before.) And our experience is in no sense isolated from the broader academic and professional literature. As we have already noted in Section 8.2.1.2.4, Magnus and Morgan (1999) presented strong empirical evidence for the importance of tacit knowledge in econometric modelling; and on the basis of such evidence, Mingers (2006a) rejected claims to 'researcher-independent objectivity' in the wider field of statistical modelling. Similarly, in Section 8.2.4.1.1 above, we quoted Robinson's (2008b) conclusion to the effect that no simulation model is 'absolutely right', because any conceptual model is dependent on the particular preferences and perceptions of the people involved in the study.

We would thus regard OR modelling as a craft as much as a science (cf. Ravetz, 1971). But – *contra* Ormerod (2006a) as quoted in Section 4.2 above, for instance – that is not to say that pragmatic considerations should necessarily be dominant. Instead, we have argued in this thesis that theory development ought not to be neglected in OR modelling – particularly, in terms of gaining a better understanding of cause-and-effect relationships. (More specifically, we would now also contend that Critical Realism as a metatheory could guide not just the theory-related aspects of OR modelling, but also the practical use of such models to aid policy development and managerial problem solving). What we mean to say is that, as OR practitioners, we need craft-based skills – including a large dose of creativity – to be fully effective in our modelling efforts, whether we focus primarily on theory development or the selection of suitable problem-solving actions.

9.2 Summary findings in relation to the specific research questions

In Section 1.2 above, we listed five research questions to be explored in this thesis, all of which questions could be directly traced back to the overall research theme. We are now ready to summarise our findings in relation to each of these questions in turn.

9.2.1 The characteristics of performance measurement in public service organisations

In Section 3.1 above, we discussed five specific problems and challenges for performance measurement in public service organisations.

> First, all public service organisations of the kind considered in this thesis – see also Section 9.3 below – are organised hierarchically (Jacobs *et al.*, 2006). Therefore, when analysing the performance of a particular decision making unit (DMU), we should take account of the potential impact of decisions taken at higher or lower organisational levels.
>
> Second, all public service organisations have multiple stakeholders (Dixit, 2002; Propper & Wilson, 2003; Bird *et al.*, 2005; Vissers & Beech, 2005). The main problem is that these different stakeholders tend to have their own, differing views on the performance of the public service organisation concerned.
>
> Third, all public service organisations have multiple objectives and there is typically a lack of consensus between the different stakeholders about the choice of objectives to be measured (in terms of outputs or outcomes) and the valuations or weights to be placed on these objectives (Van Peursem *et al.*, 1995; Smith, 1995b; Stone, 2002; Williams 2003; Smith & Street, 2005).

Fourth, attempts by analysts and researchers to capture the multiplicity of outputs and outcomes in composite performance indicators are not necessarily helpful in this respect (Smith, 2002; Jacobs *et al.*, 2006; Jacobs & Goddard, 2007).

Fifth, the overall nature of decision making with respect to performance measurement in public service organisations is highly political (Blundell & Murdock, 1997; Brignall & Modell, 2000; Williams, 2003; Cavalluzzo & Ittner, 2004; Smith & Street 2005). In the UK, as in a number of other OECD countries, the rise of a set of doctrines for public sector management encapsulated in the concept of the 'New Public Management' (NPM) has led to dramatic shifts in the political constraints on performance measurement in public service organisations from the 1980's onwards (Hood, 1995 and 2007; Ferlie *et al.*, 2003; Bird *et al.*, 2005; Hodgson *et al.*, 2007; Ackroyd *et al.*, 2007). In the 2000's, the NPM approach in the UK (and particularly England) seems to have culminated in 'a system of governance of public services that combined targets with an element of terror' (Bevan & Hood, 2006).

We have analysed the importance of each of these problems and challenges for our own studies of the Maritime and Coastguard Agency and the Musculo-Skeletal Unit, respectively. As we have explained in Section 8.1 above, our studies of the two UK public service organisations differed in some important respects; namely, the precise nature of the public services that we considered; the locus of organisational decision making on which we focused; and the nature of the outcome measures that we used. With respect to the last point in particular, for performance measurement to be truly effective it would always be preferable to use the most 'fair' and comprehensive measures of the outcome of a public service possible. For the MCA studies, we were able to construct a measure of service outcome in the form of the proportion of lives saved from all those involved in life-threatening incidents. For the MSU study, on the other hand, we lacked the data to construct an outcome measure based on actual improvements in the health status of eligible patients and were forced to focus on a range of output measures, such as patient waiting times, instead.

A common finding in the MCA and MSU studies was the central role played by disputes about the proper trade-off between service productivity and service outcomes. In the MCA study, the debate between the various stakeholders crystallised around the question whether the closure of a Coastguard coordination centre would result in a trade-off between staff experience (associated with higher labour productivity) and local knowledge (associated with worse service outcomes). In the MSU study, the key point at issue was how best to resolve the expected trade-off between patient numbers (associated with higher labour productivity) and waiting times (associated with worse service outcomes). In both cases, these disputes had strong political overtones, as predicted more generally by Williams (2003). This makes it very hard for OR practitioners to keep all relevant stakeholders convinced of their scientific objectivity in trying to gain this better understanding – remembering that, according to the tenets of Critical Realism, social scientists can generally only interpret other people's interpretations of external reality.

9.2.2 The application of OR models in support of performance measurement in public service organisations

In Section 5.3 above, we discussed four different categories of OR modelling approaches.

First, models that are based on the application of problem structuring methods such as Cognitive Mapping (SODA or Journey Making) or Soft Systems Methodology. These would all be included in Pidd's (2003) list of soft OR methods; and they would fall under Sanderson & Gruen's (2006) first heading of "models for clarifying complex decisions".

Second, models (usually, but not always, deterministic) that involve the application of some kind of optimisation method, typically of a mathematical nature (e.g. mathematical programming or heuristic search methods). These would all be included in Pidd's list of hard OR methods; and they would fall under Sanderson & Gruen's second heading of "models for planning and allocating resources".

Third, models (usually, but not always, stochastic) that involve the application of some kind of simulation, or other non-optimising analytical, method. Some of these – such as Markov models, Queueing Theory or Discrete Event Simulation – would normally be included in Pidd's list of hard OR methods; and others – such as System Dynamics and Bayesian Belief Nets – in his list of soft OR methods. All of these models, including arguably BBN's, would appear to fall under Sanderson & Gruen's third heading of "models for evaluating effects of changes in systems".

Fourth, models for measuring performance; or, more particularly, models for measuring performance differentials between decision-making units that are broadly comparable otherwise. Data Envelopment Analysis and Stochastic Frontier Analysis are important examples of these kinds of models. Their specific role would not seem to be adequately captured by Pidd's hard-soft dichotomy (although DEA may be seen as 'softer' than SFA as it seems to involve a greater degree of interpretive modelling). Also, these kinds of models are not considered by Sanderson & Gruen, although – if one had to include them in any of their three categories – the third heading of "models for evaluating effects of changes in systems" would appear to fit them best.

We have already indicated in Sections 2.1 and 5.3 that DEA and SFA models – as specific examples of the fourth category of OR modelling approaches listed above – play a central role in supporting performance measurement. And this is particularly true for the public sector, where the value of the outputs and outcomes produced must be valued somehow, despite the fact that 'willingness to pay' on the part of service users may not be a feasible indicator for their quality (Williams, 1993; Anthony & Young, 2002). For example, in their 2005 textbook on efficiency and productivity analysis, Coelli *et al.* focus on a limited range of methods; namely, index numbers and DEA, followed by econometric estimation (through regression analysis), in general, and SFA, in particular. Similarly, DEA and SFA constitute the main approaches discussed by Jacobs *et al.* in their 2006 textbook on efficiency measurement in public service organisations providing health care.

We have also already argued that models in the third category of OR modelling approaches listed above are essentially complementary to approaches such as SFA and DEA that are more directly concerned with measuring differences in organisational performance. As explained in Section 8.2.4.2 above (partly recalling our discussion in Section 4.1), there are (at least) eight specific goals which managers of public service organisations may wish to pursue; namely to: evaluate; control; budget; motivate; promote; celebrate; learn; and improve. But to foster improvement could be regarded as the core purpose behind the other seven (Behn, 2003). However, for improvement to be feasible, one needs to know 'what works, for whom, when, in what circumstances and why' (Ferlie *et al.*, 2003; Hodgson *et al.*, 2007). And this goes for managers of public service organisations just as much as for managers of any other kind of organisation – except that public service managers face specific organisational problems and challenges that make it particularly hard for them to measure performance effectively, as we have already stated in the summary answer to our first research question. In short, for performance measurement to be useful, we need some tool that enables us to understand what actually causes good (or bad) performance. Therefore, the main reason why DES and BBN models – as specific examples of the third category of OR modelling approaches listed above – can play an important role in support of public sector performance measurement is that they allow us to examine the cause-and-effect relationships inside the black box of the process transforming inputs into outputs and outcomes.

In the four papers under consideration, we have focused on OR models in the third and fourth categories listed above, as the relevance of models in the first and second categories seemed to be more limited in the present context. Rather than explicitly applying the methods included in the first category above, problem structuring issues were dealt with in the process of developing the initial regression models for the

MCA studies and the initial simulation models for the MSU study. The application of mathematical optimisation methods as included in the second category above, on the other hand, did not seem to be directly relevant to the particular issues that we were faced with in the MCA studies. With regard to the MSU study, any attempt at optimisation was on the basis of trial-and-error (based on the results of simulated experiments) rather than the application of formal mathematical methods. In consequence, we are not able to comment further on the application of problem structuring methods or optimisation models in support of public sector performance measurement on the basis of the MCA and MSU studies.

We were able, on the other hand, to examine the applicability of general principles of OR modelling in the MCA and MSU studies; and we have already discussed our conclusions on this question in Section 9.1.1. To summarise, we found – not unexpectedly – that most of these general principles could not be applied unquestioningly in the specific contexts provided by our own empirical studies. In particular, we concluded that: the applicability of the 'model simple' principle depends on the particular purposes that an OR study is aimed at; the universal nature of the 'start small and add' principle is not guaranteed in practice; the 'divide and conquer' principle implies that potential problems of multimethodology should be confronted; the 'use metaphors etc.' principle may be interpreted as indicating a need for model transparency; the applicability of the 'do not fall in love with data' principle may be affected by problems with data availability; and, perhaps most importantly, the 'model building may feel like muddling through' principle indicates the importance of tacit knowledge in OR modelling and the lack of researcher-independent objectivity, suggesting that OR modelling is a craft as much as a science.

9.2.3 The role of a multimethodology approach to OR modelling

A multimethodology approach to OR modelling has a number of potential advantages (Mingers & Brocklesby, 1997; Mingers, 2003):

- real-world problem situations are highly complex and multidimensional, and different research methodologies allow an investigator to focus on different aspects of the situation;
- different methodologies may be used in different phases of modelling process;
- the combination of different methodologies can improve the reliability of research results through triangulation;
- and, in any case, different methodologies are already widely combined in practice – usually within a kind of paradigm that is based on Pragmatism (cf. Tashakkori & Teddlie, 1998; Ormerod, 2006a) or a Postmodernist perspective (cf. Schultz & Hatch, 1996; Spender, 1998).

However, these advantages are not automatic. There are potential problems in having to bridge the 'ontological divides' associated with different scientific paradigms, to which the concept of triangulation does not necessarily provide a satisfactory answer (Bryman, 2007). The potential barriers to the successful application of multimethodology research in general, and multiparadigm research in particular, can be summarised as follows (Mingers & Brocklesby, *op. cit.*):

- cognitive – the problem of an individual researcher moving easily from one scientific paradigm to another;
- cultural – the extent to which organisational and academic cultures militate against multi-paradigm work;
- philosophical – 'paradigm incommensurability', which would occur if methodologies are thought to belong to different paradigms that are considered to be incompatible with each other.

In the MSU-DES study, there were no issues relating to multimethodology. The only methodology used was that of discrete-event simulation modelling, which is usually regarded as belonging to hard OR. We did not apply any soft OR methodologies in the form of problem-structuring methods in preparation of the simulation modelling, as the general outline of the client's problem was already reasonably clear from similar studies done by OR practitioners elsewhere.

In the MCA studies, on the other hand, a multimethodology approach was developed over time. Given the highly aggregate nature of the annual Coastguard statistics that were initially available, we focused first on developing an LRA model in order to identify statistically the main explanatory factors affecting the annual proportion of lives saved in each Coastguard district. Our focus then shifted towards efficiency analysis based on the closely-related methodology of SFA. Given this new focus, it was then a logical step to compare the SFA results with those from another methodology for efficiency analysis, namely DEA. A much bigger methodological step was taken subsequently, when it was decided to develop a BBN. If we had treated the different methodologies in relative isolation from each other, then multimethodology might not have brought any great benefits. But the results from the application of different methodologies in the MCA papers are directly complementary to each other. For instance, the construction of the (micro-level) BBN was directly informed by the results from the earlier (macro-level) regression and efficiency analysis and, in turn, the structure of the BBN sheds more light on the meaning of those earlier results.

As far as the potential barriers to multimethodology research are concerned, the cognitive problem of using methodologies with which not everyone was very familiar was resolved by working as a team of researchers with complementary skills. However, the key requirement for a successful outcome was that all members of the research team had to be willing to gain at least a basic familiarity with methodologies with which they had previously been unfamiliar. On the other hand, the cultural problem of switching between different research methodologies and, if appropriate, scientific paradigms is no real problem at all in a Department of Management Science that explicitly prides itself on its capability to accomplish this as a group of academic researchers.

The philosophical problem of paradigm incommensurability only became somewhat of an issue with the development of a BBN. The initial Logistic Regression Analysis and also the subsequent efficiency analysis (SFA and DEA) were primarily based on objective assessments of performance, belonging to the realm of hard OR. Unlike regression analysis – which is founded on a 'classical', objectivist approach to statistics – BBN's are founded on a 'Bayesian', subjectivist approach to statistics. But there were a number of reasons why any problem of paradigm incommensurability in the MCA studies ultimately seemed to be relatively minor. A key reason was that we followed a well-established and rigorous process in conducting the interviews with the MCA experts that formed the main input into the construction of the BBN. This process was designed to minimise as much as possible the effects of various biases that are known to affect expert judgements in such situations. Moreover, we could exercise some level of empirical control through our statistical analysis of SAR incident data. Perhaps the main safeguard, however, was that the interviews with the MCA experts did not touch directly on controversial issues such as centre closures, but were concerned with the much more general question of what are the main risk factors determining the outcome of an individual SAR operation.

9.2.4 The kind of research philosophy to be adopted

Multimethodology may fail because of paradigm incommensurability, if methodologies are thought to belong to different scientific paradigms that are considered to be incompatible with each other. This would appear to be a particular problem in the general area of social science (or, more particularly, organisation and management) research, of which our research on performance measurement in public service organisations would form but a specific example (Burrell & Morgan, 1979). Specifically, there is a widespread acceptance – including among members of the wider community of OR practitioners (e.g. see Roy, 1993; and Pidd, 2003) – that there are crucial differences between a paradigm incorporating a realist ontology together with an empiricist or positivist epistemology at one possible end of the spectrum (akin to hard OR modelling), and a paradigm incorporating a non-realist ontology together with a constructivist or interpretivist epistemology at the other end (akin to soft OR modelling). In Section 7.1 we have already concluded that neither of these popular paradigms is, on its own, wholly suitable for our specific area of research. The first alternative incorporates an empiricist/positivist epistemology that tends to ignore the different perspectives – on the desirability of potential outcomes but also the probability of their occurrence as well as their essential nature – which different stakeholders in the public sector may hold and that, moreover, tends to deny the roles that these stakeholders may play in shaping potential outcomes. The second alternative incorporates a non-realist ontology that cannot cope very well with vital issues in public sector performance measurement such as the problem of unintended consequences or the need for public accountability.

There would appear to be three different ways of dealing with paradigm incommensurability.

First, one could take a pragmatic approach and largely ignore the problem (Tashakkori & Teddlie, 1998; Aram & Salipante, 2003; Morgan, 2007). According to this line of reasoning, taking a pragmatic approach means that we should largely ignore any problem of mixing realist and non-realist ontological perspectives. But that implies that we would be ignoring the issue that tends to give rise to paradigm incommensurability in the first place.

Second, one could apply a contingency approach to model validation, possibly combined with triangulation through reciprocal validation of the results from different methodologies. In this line of argument, the hard aspects of the model would be validated according to empiricist/positivist criteria and the soft aspects according to constructivist/empiricist criteria. (References in the more general social science literature include: Lincoln & Guba, 1985; Guba & Lincoln, 1994; Sale & Brazil, 2004; Johnson *et al.*, 2006; and in the OR literature: Oral & Kettani, 1993; Roy, 1993; Pidd, 2003). This could then even lead to a process of triangulation by which the results from different methodologies would be reciprocally validated. In the wider area of organisation and management research, this form of triangulation appears to be increasingly regarded as a key advantage of multimethodology research, if not its main purpose (Gioia & Pitre, 1990; Lewis & Grimes, 1999; Bryman, 2001; Johnson *et al.*, 2007). However, rather than trying to bridge the ontological divide between hard, mainly quantitative, and soft, mainly qualitative, methodologies by trying to establish a unified position on the nature of reality, triangulation is regarded as a means towards 'forging a negotiated account' between different ontological positions (Brannen, 2005; Bryman, 2007). But that implies that, although the problem of paradigm incommensurability is no longer ignored, it has still not been fully resolved.

Third, one could adopt a new philosophical position that can incorporate both hard, mainly quantitative, and soft, mainly qualitative, research methodologies – and thereby gaining a new perspective on multimethodology. The viability of such an approach depends crucially on the answer to a key question; namely, whether or not particular research methodologies are inextricably associated with specific ontological and/or epistemological axioms. To be more specific, are hard, mainly quantitative, methodologies inextricably connected to a realist ontology and an empiricist or positivist epistemology? Or are soft, mainly qualitative, methods inextricably linked to a non-realist ontology and a constructivist or interpretivist epistemology? The correct answer would appear to be 'no' (cf. the arguments made by Halfpenny, 2005, as quoted in Brannen, 2005). Specifically, Critical Realism represents a philosophical position founded on a realist ontology that can nevertheless be combined with a constructivist or interpretivist type of epistemology, as we have explained in some detail in Section 7.2.3 above – thus resolving the problem of paradigm incommensurability for our specific area of research.

The adoption of Critical Realism as our preferred research philosophy would enable us to avoid the problems associated with the application of either an empiricist or positivist epistemology or a non-realist ontology to our specific research area of performance measurement in public service organisations. Moreover, Critical

Realism contains the idea of 'Critical Methodological Pluralism' (Danermark *et al.*, 2002), through which we could satisfactorily resolve any problem of combining hard, mainly quantitative, and soft, mainly qualitative, research methodologies in our studies. (As long as we are able to treat all these different methodologies as consistent with a realist perspective on ontology.) But it is important to stress that we have not explicitly tried to take a Critical Realist position in any of our four papers currently under consideration. Instead, we tended to take the pragmatic view of largely neglecting ontological concerns. As already explained in our answer to the third research question, in our own studies any problem of paradigm incommensurability ultimately seemed to be relatively minor. Without necessarily saying this in so many words, we always stuck to a realist perspective on ontology and tended link this to a generally empiricist or positivist stance on epistemology. Our realisation of the problems associated with the application of an empiricist or positivist epistemology in our specific area of research only developed as we were working on the construction of a BBN for the MCA case.

On the basis of the arguments developed in this thesis, we conclude that OR practitioners (particularly those favouring a multimethodology approach) should clarify – both to themselves and to other stakeholders (for, as OR practitioners, we should expect to be seen as stakeholders, however much we might protest our innocence of that charge) – their overarching philosophical position, including their particular stance on ontological and epistemological questions. Applying a range of methodologies can significantly enhance our ability to look at a problem from different angles and thereby to enhance our understanding, as long as we do not ignore problems of paradigm incommensurability (for example, by trying to mix-and-match realist and non-realist perspectives of the world). In research that seeks to combine different OR models in support of performance measurement in public service organisations, we propose the adoption of Critical Realism as a good (though not necessarily the only) way for resolving such problems.

9.2.5 The type of model validation criteria to be applied

In our answer to the fourth research question, we mentioned the contingency approach to model validation – which would suggest different validation criteria for soft, mainly qualitative, OR models, as opposed to hard, mainly quantitative, ones – and argued that it could not fully resolve the problem of paradigm incommensurability. Rather than taking a contingency approach, it would seem preferable to find a set of, suitably generic, validation criteria that could be applied effectively to a relatively wide range of different OR models. For if we use the supposed lack of universal validation criteria as a pragmatic excuse to select whatever criteria would suit our own purposes as OR practitioners best, then how could we convince other stakeholders of our scientific objectivity? Adopting a Critical Realist position will enable us to achieve more unified approach to model validation in a way that fits in with our own basic tendency as OR practitioners to stick to a realist perspective on ontology.

We propose that those OR models that are founded – like ours – on a realist ontology should be validated according to two groups of validation criteria. The group of primary criteria would consist of the four classical validation criteria outlined by Yin (2003) and many other social science researchers (as explained in Section 7.2.2): construct validity, internal validity, external validity, and reliability. The group of secondary criteria would consist of a variety of pragmatic criteria to evaluate the usefulness of the model under consideration.

The group of primary criteria is based on a combination of realist and axiomatic perspectives on model validation. More specifically, the four classical validation criteria should be interpreted in light of the Critical Realist position, so as to make them equally applicable to hard, mainly quantitative (i.e. 'extensive') and soft, mainly qualitative (i.e. 'intensive') methodologies – as follows.

> First, construct validity refers to establishing the correct operational measures for the concepts being studied. In a Critical Realist approach to OR

modelling, concepts take the form of conceptual abstractions relating not just to the 'empirical' domain of reality, but also to the 'actual' domain and the 'real' domain. In other words, the question of construct validity goes well beyond quantitative measurement issues in the empirical domain and is equally (if not more so) relevant to the definition of qualitative constructs in the real domain (such as the generative mechanisms that are the ultimate focus of Critical Realist theorising).

Second, internal validity refers to establishing a causal relationship, whereby certain conditions are shown to lead to other conditions, as distinguished from spurious relationships. The main aim of the Critical Realist position is to gain knowledge about causal mechanisms linking the empirical, actual and real domains (but analysed as tendencies, not universal empirical regularities). Interpreted in this way, internal validity is the key criterion for evaluating an OR model, whatever the (mix of) research methodologies applied. In fact, as we have seen, Critical Realism advocates the application of complementary – hard/extensive and soft/intensive – methodologies (under the label of Critical Methodological Pluralism) to gain such knowledge.

Third, external validity refers to establishing the domain to which a study's findings can be generalised. According to the Critical Realist position, external validation is not based on checking whether empirical phenomena predicted by the model are generally occurring. Instead, generalisation is interpreted as using 'retroductive' inference (Danermark *et al.*, 2002) to find fundamental structures and causal mechanisms operating in the real domain, as well as establishing the region of time-space for which these mechanisms are actualised in the form of 'demi-regularities' (Lawson, 1997). Again, both extensive and intensive methodologies should be used for this task. Another requirement explicitly imposed by Critical Realism (but not always expressed as forcefully by other approaches) is to seek out competing models and then to try eliminate as many of these as possible on empirical grounds. A particularly strong model is one that could account for all of the empirical phenomena explained by rival models but, in addition, could explain some important phenomena that other models leave unexplained.

Fourth, reliability refers to demonstrating that the operations of a study can be repeated, with the same results. As explained by Tsang & Kwan (1999), although the general infeasibility of artificially creating closed systems in social science research – including organisation and management studies and, therefore, also OR modelling – makes replication difficult, it is not impossible in principle. In fact, Tsang & Kwan refer to six different types of replication that could be tried, each of which would apply to both extensive and intensive methodologies.

The group of secondary criteria is based on a combination of constructivist, instrumentalist and axiomatic perspectives on model validation. The main focus of these criteria is the usefulness of the OR model. However, usefulness can be interpreted in many different ways – as demonstrated by Oral & Kettani (1993), Smith (1993), and many other OR practitioners since. So, as the usefulness of an OR model cannot be divorced from the uses to which it may be put (Miser, 1993; Pidd, 2003), it is impossible to give a definitive and exhaustive list of headings for this group. But we would argue that all of these secondary criteria should be understood on the basis of Fleetwood & Ackroyd's (2000) contention that Critical Realism as a metatheory should guide not just the theory-related aspects of OR modelling, but also the practical use of such models for policy development and managerial intervention.

In our own MCA and MSU studies, we applied a range of criteria of both a classical (primary) and pragmatic (secondary) nature. Overall, we stuck to a realist ontology and we tended to look for open-box rather than black-box validation. In the MCA studies, our main focus has been on the classical validation criteria of construct validity, internal and external validity, and reliability. Our range of validity checks on the LRA, SFA and DEA models represented some form of grey-box validation in which we were mainly concerned with construct validity, external validity and reliability, but where the internal structure of the model was only considered at the macro level. With the development of the micro-level Bayesian Belief Net (BBN) we moved much further towards open-box validation, although the reliability of the research process remained an issue of particular importance from a validation perspective. As the key purpose of the MCA studies was to expose managerial decisions about the organisation of the Coastguard service to public accountability, the usefulness of our MCA models was best served by trying to apply the classical validation criteria as effectively as possible – but with a particular emphasis on making the internal structure of our models more transparent to non-specialists in OR. In the MSU study, we have tended to take care of the classical validation criteria by applying Robinson's (2004) guidelines for the validation of Discrete Event Simulation models. The two forms of validation that turned out to be most problematic were black-box validation and solution validation, as explained in more detail in Section 7.4 above. As a key purpose of the MSU study was to facilitate organisational learning and thus to help our clients make a case for further investments in the service, that particular purpose constituted the main basis for any usefulness criteria for our sequence of simulation models. Figure 7 (p. 170) in the MSU-DES paper is meant to summarise our views on this issue.

9.3 Review of the boundaries of this research

We shall now reconsider an issue that we first raised in Section 1.3 above; namely, the boundaries of the research discussed in this thesis. As we have stated earlier, while the overall research theme of this thesis is very broad, the four papers that form our own contribution to the empirical basis for this thesis are, inevitably, rather more limited in their scope. So how applicable are our conclusions to other empirical studies relating to OR modelling for public sector performance measurement?

The evidence base for our conclusions consists of the findings from our own empirical research in conjunction with the evidence from the academic and professional literature that we have consulted. In this thesis (including the four papers under consideration), we have cited a wide range of external sources – many referring to theoretical frameworks that seem appropriate for interpreting our own empirical findings; and many others reporting on the results from empirical studies that may in some way be compared or contrasted to our own. In particular, our answers to the fourth and fifth questions from our list of research questions have been primarily explored through our detailed analysis of relevant elements in the wide-ranging academic literature on research philosophy and model validation in Chapter 7 of this thesis. For our conclusions on the appropriate research philosophy and model validation criteria for a multimethodology approach to OR modelling we may, therefore, fairly claim a relatively high degree of generalisability (i.e. maintaining its external validity across a wide range of OR modelling approaches and public sector organisations), albeit mainly on a theoretical level.

Similarly, our answer to the first research question is mainly based on a careful study of the academic and professional literature on performance measurement in public service organisations, the outcome of which study is reported in Chapters 2 and 3. More particularly, we have explained the main points from the relevant literature in Sections 2.1, 3.1 and 3.4; and we have demonstrated in Sections 2.2 – 2.5, 3.2 – 3.3 and 3.5 – 3.6 exactly how these points manifested themselves in the MCA and MSU studies. Therefore, our conclusions on the particular characteristics of performance measurement in public service organisations are meant to be generally applicable, not least because they are based on theoretical and empirical evidence from a wide range of external sources in addition to our own studies.

For our answers to the second and third research questions, on the other hand, we have relied more heavily on the case study material from our own MCA and MSU studies to complement the existing evidence from the OR literature. Although one should always be very wary about making inductive generalisations from a small number of practical examples, we are entitled to claim that: (a) case study research is a valuable methodology in its own right (Yin, 2003), particularly when one wishes to understand the complex ways in which fundamental causal mechanisms may interact in specific organisational settings (Danermark *et al.*, 2002); and (b) the external validity of our case study findings has been enhanced by their comparison – in Chapter 8 – both with each other and with the results from relevant studies conducted by other researchers. Specifically, the OR models that we have applied in our own studies are particular examples of modelling approaches in – what we have termed in Section 5.3 above – 'category 4' models for measuring performance and 'category 3' models involving the application of some kind of simulation, or other non-optimising analytical, method. In other words, we have not used either problem structuring ('category 1') or optimisation ('category 2') models in the four papers under consideration, although we have referred to many general aspects of OR modelling in Chapters 4 and 5, and (with specific reference to multimethodology) Chapter 6. Also, the public service organisations in the MCA and MSU studies are particular examples relating to emergency services and health care services, respectively. But, as explained in the previous paragraph, we have demonstrated in Chapters 2 and 3 (in particular, Sections 2.2 – 2.5, 3.2 – 3.3 and 3.5 – 3.6) how our empirical studies reflected the wider characteristics of performance measurement in public service organisations. Nevertheless, with respect to the second and third research questions, we should make the general proviso that the validity of our answers may not be wholly independent from the extent to which the type of public sector organisation considered and/or the kind of OR models applied resemble our own examples.

9.4 Final comments and questions for future research

We shall complete this chapter with some final comments on our MCA and MSU studies, respectively, and a number of questions for future research. In the MCA-LRA paper, we used our LRA model to analyse a set of panel data, consisting of aggregate Coastguard statistics for the years 1995-1998 (with the 1999 data used for a prediction test); that is, before the proposed closures of some coordination centres had been put into effect. Since then, of course, Coastguard statistics have become available for subsequent years. This has enabled us to re-estimate and test our LRA and SFA models on the basis of a much larger data set, including the periods immediately before, during and after the centre closures. We hope to publish the detailed results of our updated analysis in the near future. Briefly, the most important conclusions are that, while the importance of the set of explanatory variables identified earlier is reconfirmed, there is some statistical evidence of a 'structural break' in the regression model around the time of the closures (2000-2001); also, the effects of centre closures on the SAR performance of the remaining groups of coordination centres appear to be mixed, rather than uniformly negative or positive. Finally, our recent incorporation of a (as yet very limited) set of data on SAR incidents involving suicide or criminal acts appears to make relatively little difference to the results from our analysis.

As we have already mentioned in Section 8.2.2.3.5, our DEA models did not yield many additional insights beyond confirming some, but not all, of the earlier SFA results. Therefore, we are not planning to extend the analysis in the MCA-DEA paper. In contrast, we do want to develop the BBN model reported in the MCA-BBN paper. As explained in Section 8.2.3.2 above, we intend to collect and analyse statistically a larger set of individual incident records in order to strengthen the BBN's empirical basis and to triangulate our findings from the macro and micro levels of analysis. The key question for our future MCA studies is how best to piece together the empirical evidence from different sources, including aggregate Coastguard statistics, individual incident records and interviews with various domain experts, within a common theoretical framework.

With respect to the research reported in our MSU-DES paper, we have already mentioned two issues – in Sections 8.2.4.1.1 and 8.2.4.1.2, respectively – that merit further attention. The first issue is associated with the nature of our DES models, which tended to become more complex, in line with the questions asked by our MSU clients. Developing and maintaining more complex simulation models typically demands ever more effort (and not just from the OR practitioners, but also from clients and, possibly, other stakeholders). This raises the question whether and, if so, how we can use the knowledge and understanding thus gained in order to distinguish more effectively between necessary and irrelevant details and to construct more simple DES models as a result. In other words, the question is about the extent to which we could apply some form of the 'general-to-specific' modelling approach – as applied in modern econometrics (Campos *et al.*, 2005); see Section 9.1.2 above – to the construction of simulation models.

The second issue arising from our MSU-DES paper concerns the likelihood of feedback effects in health care service systems and how best to incorporate these into our simulation models. Since feedback effects are typically modelled using SD (see Section 8.2.4.1.3 above) – as opposed to the use of DES models in our MSU study – this raises the question whether and, if so, how DES and SD models should be combined in practice. Although this has already been a matter for considerable academic debate (Lane, 2000; Brailsford & Hilton, 2001; Morecroft & Robinson, 2005), it appears that more empirical evidence is still needed (Tako & Robinson, 2008), and future studies of the MSU may provide a suitable context for this.

The question of how SD models could be applied to future MSU studies suggests a broader issue; namely, the use of a wider range of OR modelling approaches in support of performance measurement in public sector organisations. As explained in Section 9.3 above, so far, we have applied a limited range of OR modelling approaches. More particularly, we have not used problem structuring or optimisation models, which raises the question what the advantages and disadvantages might be of including either of these categories of OR models into the multimethodology approach that we have discussed in this thesis. Also, we have only applied our OR models in the context of a couple of specific public service organisations; this would suggest the usefulness of conducting similar studies in a wider range of public sector organisations, in order to gain a better understanding of any contingency factors relating to the precise nature of the public services and the particular locus of organisational decision making considered in a study.

10 BIBLIOGRAPHY

Abi-Zeid, I. and Frost, J. R. (2005). SARPlan: A decision support system for Canadian Search and Rescue Operations, *European Journal of Operational Research 162 (3)*, pp. 630–653.

Ackroyd S., Hughes J.A. and Soothill K. (1989), Public sector services and their management, *Journal of Management Studies 26 (6)*, pp. 603-619.

Ackroyd S., Kirkpatrick I. and Walker R.M. (2007), Public management reform in the UK and its consequences for professional organization: a comparative analysis, *Public Administration 85 (1)*, pp. 9-26.

Allen R., Athanassopoulos A., Dyson R.G. and Thanassoulis E. (1997), Weights restrictions and value judgements in Data Envelopment Analysis: evolution, development and future directions, *Annals of Operations Research 73*, pp. 13-34.

Anthony R.N. and Young D. (2002), *Management Control in Nonprofit Organizations*, 7th edition, McGraw-Hill / Irwin.

Aram J.D. and Salipante P.F. (2003), Bridging scholarship in management: epistemological reflections, *British Journal of Management 14*, pp. 189-205.

Archer M., Bhaskar R., Collier A., Lawson T. and Norrie A. (eds.) (1998), *Critical Realism: Essential Readings*, London: Routledge.

Bacharach S.B. (1989), Organizational theories: some criteria for evaluation, *The Academy of Management Review 14 (4)*, pp. 496-515.

Bauer P.W., Berger A.N., Ferrier G.D. and Humphrey D.B. (1998), Consistency conditions for regulatory analysis of financial institutions: a comparison of frontier efficiency methods, *Journal of Econometrics and Business 50 (2)*, pp. 85-114.

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

Bevan G. and Hood C. (2006), What measured is what matters: targets and gaming in the English public health care system, *Public Administration 84 (3)*, pp. 517-538.

Bhaskar R. (1978), *A Realist Theory of Science*, Hassocks: Harvester Press.

Bird S.M., Cox D., Farewell V.T., Goldstein H., Holt T. and Smith P.C. (2005), Performance indicators: good, bad, and ugly, *Journal of the Royal Statistical Society Series A 168, Part 1*, pp. 1-27.

Blundell B. and Murdock A. (1997), *Managing in the public sector*, Oxford: Butterworth-Heinemann.

Boland T. and Fowler A. (2000), A systems perspective of performance management in public sector organisations, *The International Journal of Public Sector Management 13 (5)*, pp. 417-446.

Brailsford S. and Hilton N. (2001), A comparison of discrete event simulation and system dynamics for modelling health care systems, in: Riley J. (ed.), *Proceedings of ORAHS 2000*, Glasgow, pp. 18-39.

Brannen J. (2005), Mixed methods research: a discussion paper, *Methods Review Paper NCRM/005, ESRC National Centre for Research Methods*.

Brignall S. and Modell S. (2000), An institutional perspective on performance measurement and management in the 'new public sector', *Management Accounting Research 11*, pp. 281-306.

Bryman A. (2001), *Social Research Methods*, Oxford: Oxford University Press.

Bryman A. (2007), Barriers to integrating quantitative and qualitative research, *Journal of Mixed Methods Research 1 (1)*, pp. 8-22.

Buchanan J.T., Henig E.J. and Henig M.I. (1998), Objectivity and subjectivity in the decision making process, *Annals of Operations Research 80*, pp. 333-345.

Burrell G. and Morgan G. (1979), *Sociological Paradigms and Organisational Analysis: Elements of the Sociology of Corporate Life*, London: Heinemann.

Campos J., Ericsson N.R. and Hendry D.F. (eds.) (2005), *General-To-Specific Modelling*, Cheltenham: Edward Elgar.

Cavalluzzo K.S. and Ittner C.D. (2004), Implementing performance measurement innovations: evidence from government, *Accounting, Organizations and Society 29*, pp. 243-267.

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

Charnes A., Cooper W.W., Huang Z.M. and Sun, D.B. (1990), Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks, *Journal of Econometrics 46*, pp. 73-91.

Checkland P. (2006), Reply to Eden and Ackermann: Any future for problem structuring methods?, *Journal of the Operational Research Society 57*, pp. 769-771.

CHI (2003), *Ambulance Trusts*, London: Commission for Health Improvement.

Coelli T.J., Prasada Rao D.S., O'Donnell C.J. and Battese G.E. (2005), *An Introduction to Efficiency and Productivity Analysis*, 2nd edition, New York, NY: Springer.

Collier P.M. (2006), In search of purpose and priorities: police performance indicators in England and Wales, *Public Money & Management, June 2006*, pp. 165-172.

Cook W.D., Kress M. and Seiford L.M. (1992), Prioritization models for frontier decision making units in DEA, *European Journal of Operational Research 59*, pp. 319-323.

Cooper K., Brailsford S.C. and Davies R. (2007), Choice of modelling technique for evaluating health care interventions, *Journal of the Operational Research Society 58 (2)*, pp. 168-176.

Cooper W.W., Seiford L.M. and Tone K. (2006), *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*, 2nd edition, New York, NY: Springer.

Danermark B., Ekström M, Jakobsen L. and Karlsson J.Ch. (2002), *Explaining society – critical realism in the social sciences*, Abingdon: 2002.

Davies R. and Davies H.T.O. (1994), Modelling patient flows and resource provision in health systems, *Omega– The International Journal of Management Science 22*, pp. 123-131.

Deetz S. (1996), Describing differences in approaches to organization science: rethinking Burrell and Morgan and their legacy, Organization Science 7 (2), pp. 191-207.

Déry R., Landry M. and Banville C. (1993), Revisiting the issue of model validation in OR: an epistemological view, *European Journal of Operational Research 66*. pp. 168-183.

Dixit A. (2002), Incentives and organizations in the public sector: an interpretative review, *The Journal of Human Resources 37 (4)*, pp. 696-727.

Donaldson J.F. (Lord) (1999), *Review of the Five Year Strategy for H.M. Coastguard*, London: Department of the Environment, Transport and the Regions.

Dyson R.G. and Thanassoulis E. (1988), Reducing weight flexibility in Data Envelopment Analysis, *Journal of the Operational Research Society 39 (6)*, pp. 563-576.

Eldabi T., Paul R.J. and Young T. (2007), Simulation modelling in health care: reviewing legacies and investigating futures, *Journal of the Operational Research Society 58 (2)*, pp. 262-270.

Eleye-Datubo, A.G., Wall, A., Saajedi A. and Wang, J. (2006), Enabling a powerful marine and offshore decision-support solution through Bayesian network technique, *Risk Analysis 26 (3)*, pp. 695–721.

Ferlie E., Hartley J. and Martin S. (2003), Changing public service organizations: current perspectives and future prospects, *British Journal of Management 14*, pp. S1-S14.

Fleetwood S. and Ackroyd S. (eds.) (2000), *Critical Realism in Action in Organisation and Management Studies*, London: Routledge.

Flood R.L. and Jackson M.C. (1991), *Creative Problem Solving – Total Systems Intervention*, Chichester: Wiley.

Foreman-Peck J. (2002), Comment on Stone M. (2002), How not to measure the efficiency of public services (and how one might), *Journal of the Royal Statistical Society Series A 165, Part 3*, pp. 425-426.

Francis G. and Holloway J. (2007), What have we learned? Themes from the literature on best-practice benchmarking, *International Journal of Management Reviews 9 (2)*, pp. 1-19.

Gioia D.A. and Pitre E. (1990), Multiparadigm perspectives on theory building, The Academy of Management Review 15 (4), pp. 584-602.

Gibbons M., Limoges C., Nowotny H., Schwartzman S., Scott P. and Trow M. (1994), *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*, London: Sage.

Goles T. and Hirschheim R. (2000), The paradigm is dead, the paradigm is dead ... long live the paradigm: the legacy of Burrell and Morgan, *Omega – The International Journal of Management Science 28*, pp. 249-268.

Guba E.G. and Lincoln Y.S. (1994), Competing paradigms in qualitative research, in *N.K. Denzin & Y.S. Lincoln (eds.), Handbook of Qualitative Research*, Thousand Oaks, CA: Sage.

Guedes Soares, C. and Texeira, A.P. (2001), Risk assessment in maritime transportation, *Reliability Engineering and System Safety 74 (3)*, pp. 299–309.

Halfpenny P. (2005), The reflection between quantitative and qualitative research, *Paper Presented at the Mixed Methods Workshop, ESRC Research Methods Programme, October*, Manchester.

Hall C.S. and Lindzey G. (1957), *Theories of Personality*, New York: Wiley.

Hannesson R. (2005), Time-inconsistency in DEA: a study of Norwegian factory trawlers, *SNF Working Paper No. 22/05, Institute for Research in Economics and Business Administration,* Bergen, Norway.

Harré R. (1970), *The Principles of Scientific Thinking*, London: Macmillan.

Heath G. and Radcliffe J. (2007), Performance measurement and the English ambulance service, *Public Money & Management, June 2007*, pp. 223-227.

Hodgson L., Farrell C.M. and Connolly M. (2007), Improving UK public services: a review of the evidence, *Public Administration 85 (2)*, pp. 355-382.

Hood C. (1995), The "new public management" in the 1980s: variations on a theme, *Accounting, Organizations and Society 20 (2/3)*, pp. 93-109.

Hood C. (2007), Public service management by numbers: Why does it vary? Where has it come from? What are the gaps and the puzzles?, *Public Money & Management, April 2007*, pp. 95-102.

HoC Transport Committee (2005), *SAR Report*, London: House of Commons.

Hosmer D.W. and Lemeshow S. (2000), *Applied Logistic Regression*, 2nd edition, Chichester: Wiley.

Howick S., Eden C, Ackermann F. and Williams T. (2008), Building confidence in models for multiple audiences: the modelling cascade, *European Journal of Operational Research 186 (3)*, pp. 1068-1083.

Jacobs R. and Goddard M. (2007), How do performance indicators add up? An examination of composite indicators in public services, *Public Money & Management, April 2007*, pp. 103-110.

Jacobs R., Smith P.C. and Street A. (2006), *Measuring Efficiency in Health Care – Analytical Techniques and Health Policy*, Cambridge: Cambridge University Press.

Johnsen Å, (2005), What does 25 years of experience tell us about the state of performance measurement in public policy and management?, *Public Money & Management, January 2005*, pp. 9-17.

Johnson L.K. (2007), Decision analytics: from back office to center stage, *Balanced Scorecard Report Article, Harvard Business Online.*

Johnson R.B., Onwuegbuzie A.J. and Turner L.A. (2007), Toward a definition of mixed methods research, *Journal of Mixed Methods Research 1 (2)*, pp. 112-133.

Kaplan R.S. and Norton D.P. (1992), The balanced scorecard: measures that drive performance, *Harvard Business Review (January-February)*, pp. 71-79.

Kaplan R.S. and Norton D.P. (2004), Measuring the strategic readiness of intangible assets, *Harvard Business Review 82 (2)*, pp. 52-63.

Kennedy, P. (2008), *A Guide to Econometrics*, 6th edition, Oxford: Blackwell Publishing.

Kotiadis K. and Mingers J. (2006), Combining PSMs with hard OR methods: the philosophical and practical challenges, *Journal of the Operational Research Society 57*, pp. 856-867.

Kuhn T. (1962), *The Structure of Scientific Revolutions*, Chicago: University of Chicago Press.

Kwan K. and Tsang E.W.K. (2001), Realism and constructivism in strategy research: a critical realist response to Mir and Watson, *Strategic Management Journal 22 (12)*, pp. 1163-1168.

Lane D.C. (2000), You just don't understand me: modes of failure and success in the discourse between system dynamics and discrete event simulation, *LSEOR*

Working Paper OR.00.34, December 2000, London: London School of Economics.

Lawson T. (1997), *Economics and Reality*, London: Routledge.

Lewin, K. (1951), *Field Theory in Social Science: Selected Theoretical Papers (Edited by D. Cartwright)*, New York: Harper & Row.

Lewis M.W. and Grimes A.J. (1999), Metatriangulation: building theory from multiple paradigms, *The Academy of Management Review 24 (4)*, pp. 672-690.

Likierman A. (1993), Performance indicators: 20 early lessons from managerial use, *Public Money & Management 13 (4)*, pp. 15-22.

Lincoln Y.S. and Guba E.G. (1985), *Naturalistic Inquiry*, Newbury Park, CA: Sage.

McIntosh K. (2005), Fiddling the figures on ambulance response times will not save lives, *The Guardian Newspaper, 29 June.*

Magnus, J.R. and Morgan, M.S. (1999), *Methodology and Tacit Knowledge: Two Experiments in Econometrics*, Chichester: Wiley.

Merrick, J.R.W. and van Dorp, R. (2006), Speaking the truth in maritime risk assessment, *Risk Analysis 26 (1)*, pp. 223–237.

Micheli P., Neely A. and Kennerley M. (2005), The roles of performance measurement in the English public sector, *Paper Presented at the EGPA Conference, 31 August – 3 September*, Bern.

Mingers J. (2000a), The contribution of critical realism as an underpinning philosophy for OR/MS and systems, *Journal of the Operational Research Society 51*, pp. 1256-1270.

Mingers J. (2000b), Variety is the spice of life: combining soft and hard OR/MS methods, *International Transactions in Operational Research 7 (6)*, pp. 673-691.

Mingers J. (2003), A classification of the philosophical assumptions of management science methods, *Journal of the Operational Research Society 54*, pp. 559-570.

Mingers J. (2005), 'More dangerous than an unanswered question is an unquestioned answer': a contribution to the Ulrich debate, *Journal of the Operational Research Society 56*, pp. 468-474.

Mingers J. (2006a), A critique of statistical modelling in management science from a critical realist perspective: its role within multimethodology, *Journal of the Operational Research Society 57*, pp. 202-219.

Mingers J. (2006b), Response from the author: Intelligence and realism in OR, *Journal of the Operational Research Society 57*, pp. 1375-1379.

Mingers J. and Brocklesby J. (1997), Multimethodology: towards a framework for mixing methodologies, *Omega – The International Journal of Management Science 25 (5)*, pp. 489-509.

Miser H.J. (1993), A foundational concept of science appropriate for validation in operational research, *European Journal of Operational Research 66*, pp. 204-215.

Mitchell G. (Chairman) (1986), *Report of the OR Society's Commission on the Future Practice of OR*, Birmingham: OR Society.

Mitchell G. (1993), *The Practice of Operational Research*, Chichester: Wiley.

Morecroft J.D.W. and Robinson S. (2005), Explaining puzzling dynamics: comparing the use of system dynamics and discrete event simulation, in: Sterman J.D. *et al.* (eds.), *Proceedings of the 23rd International Conference of the System Dynamics Society*, Boston (MA): System Dynamics Society.

Morgan D.L. (2007), Paradigms lost and pragmatism regained – methodological implications of combining qualitative and quantitative methods, *Journal of Mixed Methods Research 1 (1)*, pp. 48-76.

Morgan G. and Smircich L. (1980), The case for qualitative research, *The Academy of Management Review, October 1980*, pp. 491-500.

Munro I. and Mingers J. (2002), The use of multimethodology in practice – results of a survey of practitioners, *Journal of the Operational Research Society 53*, pp. 369-378.

Norrington L., Quigley J.L., Russell A. and Van der Meer R.B. (2008), Modelling the reliability of search and rescue operations with Bayesian belief networks, *Reliability Engineering & System Safety 93 (7)*, pp. 940-949 . **(MCA-BBN)**

Oral M. and Kettani O. (1993), The facets of the modelling and validation process in operations research, *European Journal of Operational Research 66*, pp. 216-234.

Ormerod R. (2006a), The history and ideas of pragmatism, *Journal of the Operational Research Society 57*, pp. 892-909.

Ormerod R. (2006b), The OR approach to forecasting: comments on Mingers' paper, *Journal of the Operational Research Society 57*, pp. 1371-1373.

Osborne D. and Gaebler T. (1993), *Reinventing Government: How the Entrepreneurial Spirit is Transforming the Public Sector*, Reading: Addison-Wesley.

Phillips L.D. (1984), A theory of requisite decision models, *Acta Psychologica 56*, pp. 29-48.

Pidd M. (2001), The futures of OR, *Journal of the Operational Research Society 52*, pp. 1181-1190.

Pidd M. (2003), *Tools for Thinking*, 2nd edition, Chichester: Wiley.

Pidd M. (2005), The next 50 years of OR: it takes two to tango, *Journal of the Operational Research Society 56*, pp. 480-482.

Podinovski V.V. (2005), The explicit role of weight bounds in models of Data Envelopment Analysis, *Journal of the Operational Research Society 56 (12)*, pp. 1408-1418.

Podinovski V.V. (2007), Improving Data Envelopment Analysis by the use of production trade-offs, *Journal of the Operational Research Society 58 (10)*, pp. 1261-1270.

Podinovski V.V. and Thanassoulis E. (2007), Improving discrimination in Data Envelopment Analysis: some practical suggestions, *Journal of Productivity Analysis 28*, pp. 117-126.

Propper C. and Wilson D. (2003), The use and usefulness of performance measures in the public sector, *Oxford Review of Economic Policy 19 (2)*, pp. 250-267.

Proudlove N.C., Black S. and Fletcher A. (2007), OR and the challenge to improve the NHS: modelling for insight and improvement in in-patient flows, *Journal of the Operational Research Society 58 (2)*, pp. 145-158.

Raitt R. (1979), OR and science, *Journal of the Operational Research Society 30*, pp. 835-836.

Ravetz J.R. (1971), *Scientific Knowledge and its Social Problems*, Oxford: Oxford University Press.

Robinson S. (2004), *Simulation: The Practice of Model Development and Use*, Chichester: Wiley.

Robinson S. (2008a), Conceptual modelling for simulation part I: definition and requirements, *Journal of the Operational Research Society 59 (3)*, pp. 278-290.

Robinson S. (2008b), Conceptual modelling for simulation part II: a framework for conceptual modelling, *Journal of the Operational Research Society 59 (3)*, pp. 291-304.

Rosenhead J. and Mingers J. (eds.) (2001), *Rational Analysis for a Problematic World Revisited*, 2nd edition, Chichester: Wiley.

Roy B. (1993), Decision science or decision-aid science?, *European Journal of Operational Research 66*, pp. 184-203.

Sachdeva R., Williams T. and Quigley J. (2007), Mixing methodologies to enhance the implementation of healthcare operational research, *Journal of the Operational Research Society 58 (2)*, pp. 159-167.

Sale J.E.M. and Brazil K. (2004), A strategy to identify critical appraisal criteria for primary mixed-method studies, *Quality & Quantity 38*, 351-365.

Sanderson C. and Gruen R. (2006), *Analytical Models for Decision Making*, Maidenhead: Open University Press / McGraw-Hill Education.

Sanderson I. (2001), Performance management, evaluation and learning in 'modern' local government, *Public Administration 79*, pp. 297-313.

Sayer A. (1992), *Method in Social Science: A Realist Approach*, London: Routledge.

Schultz M. and Hatch M.J. (1996), Living with multiple paradigms: the case of paradigm interplay in organizational culture studies, *The Academy of Management Review 21 (2)*, pp. 529-557.

Sigurdsson J.H., Walls L. and Quigley J. (2001), Bayesian belief nets for managing expert judgement and modelling reliability, *Quality and Reliability International 17*, pp. 181-190.

Skyttner L. and Fagrell F. (1988), *Minimising human, economic and ecological losses at sea, TRITA IS-5152*, Stockholm: Stockholm University.

Smith J.H. (1993), Modelling muddles: validation beyond the numbers, *European Journal of Operational Research 66*, pp. 235-249.

Smith P.C. (1995a), On the unintended consequences of publishing performance data in the public sector, *International Journal of Public Administration 18 (2/3)*, pp. 277-310.

Smith P.C. (1995b), Performance indicators and outcome in the public sector, *Public Money & Management, October-December 1995*, pp. 13-16.

Smith P.C. (2002), Comment on Stone M. (2002), How not to measure the efficiency of public services (and how one might), *Journal of the Royal Statistical Society Series A 165, Part 3*, pp. 422-423.

Smith P.C. (2005), Performance measurement in health care: history, challenges and prospects, *Public Money & Management, August 2005*, pp. 213-220.

Smith P.C. and Street A. (2005), Measuring the efficiency of public services: the limits of analysis, *Journal of the Royal Statistical Society Series A 168, Part 2*, pp. 401-417.

Spender J.C. (1998), Pluralist epistemology and the knowledge-based theory of the firm, *Organization 5 (2)*, 233-256.

Stone M. (2002), How not to measure the efficiency of public services (and how one might), *Journal of the Royal Statistical Society Series A 165, Part 3*, pp. 405-434.

Tako A.A. and Robinson S. (2008), Comparing discrete event simulation and system dynamics: users' perceptions, *Journal of the Operational Research Society*, advance online publication 13 February 2008.

Tashakkori A. and Teddlie C. (1998), *Mixed Methodology: Combining Qualitative and Quantitative Approaches*, Thousand Oaks, CA: Sage.

Taylor K. and Dangerfield B. (2005), Modelling the feedback effects of reconfiguring health services, *Journal of the Operational Research Society 56 (6)*, pp. 659-675.

Thorpe R. and Beasley T. (2004), The characteristics of performance management research – implications and challenges, *International Journal of Productivity and Performance Management 53 (4)*, pp. 334-344.

Tranfield D. and Starkey K. (1998), The nature, social organization and promotion of management research, *British Journal of Management 9*, pp. 341-353.

Tsang E.W.K and Kwan K. (1999), Replication and theory development in organizational science: a critical realist perspective, *The Academy of Management Review 24 (4)*, pp. 759-780.

Van der Meer R.B., Quigley J. and Storbeck J.E. (2005a), Using regression analysis to model the performance of UK Coastguard centres, *Journal of the Operational Research Society 56 (6)*, pp. 630-641. **(MCA-LRA)**

Van der Meer R.B., Quigley J. and Storbeck J.E. (2005b), Using data envelopment analysis to model the performance of UK Coastguard centres, *Journal of the Operational Research Society 56 (8)*, pp. 889-901. **(MCA-DEA)**

Van der Meer R.B., Rymaszewski L.A., Findlay H. and Curran J. (2005), Using OR to support the development of an integrated musculo-skeletal service, *Journal of the Operational Research Society 56 (2)*, pp. 162-172. **(MSU-DES)**

Van Peursem K.A., Pratt M.J. and Lawrence S.R. (1995), Health management performance – a review of measures and indicators, *Accounting, Auditing & Accountability Journal 8 (5)*, pp. 34-70.

Vissers J. and Beech R. (eds.) (2005), *Health Operations Management – Patient Flow Logistics in Health Care*, Abingdon: Routledge.

Walker, W.E. (2000), POLSSS: overview and cost-effectiveness analysis, *Safety Science 35 (1–3)*, pp. 105–121.

Waters D. (1998), *A Practical Introduction to Management Science Patient*, 2nd edition, Harlow: Financial Times / Prentice Hall.

Williams D.W. (2003), Measuring Government in the Early Twentieth Century, *Public Administration Review 63 (6)*, pp. 643-659.

Wisniewski M. and Stewart D. (2004), Performance measurement for stakeholders – the case of Scottish local authorities, *The International Journal of Public Sector Management 17 (3)*, pp. 222-233.

Wong Y.H.B. and Beasley J.E. (1990), Restricting weight flexibility in Data Envelopment Analysis, *Journal of the Operational Research Society 41 (9)*, pp. 829-835.

Xie X.J., Pendergast J. and Clarke W. (2008), Increasing the power: a practical approach to goodness-of-fit test for logistic regression models with continuous predictors, *Computational Statistics and Data Analysis 52 (5)*, pp. 2703-2713.

Yin, R.K. (2003), *Case Study Research – Design and Methods*, 3rd edition, Thousand Oaks, CA: Sage.

11 USING REGRESSION ANALYSIS TO MODEL THE PERFORMANCE OF UK COASTGUARD CENTRES

12 USING DATA ENVELOPMENT ANALYSIS TO MODEL THE PERFORMANCE OF UK COASTGUARD CENTRES

13 MODELLING THE RELIABILITY OF SEARCH AND RESCUE OPERATIONS WITH BAYESIAN BELIEF NETWORKS

14 USING OR TO SUPPORT THE DEVELOPMENT OF AN INTEGRATED MUSCULO-SKELETAL SERVICE