About the 2020 Public Services Trust

The 2020 Public Services Trust is a registered charity (No.1124095), based at the RSA. It is not aligned with any political party and operates with independence and impartiality. The Trust exists to stimulate deeper understanding of the challenges facing public services in the medium term. Through research, inquiry and discourse, it aims to develop rigorous and practical solutions, capable of sustaining support across all political parties.

In December 2008, the Trust launched a major new Commission on 2020 Public Services, chaired by Sir Andrew Foster, to recommend the characteristics of a new public services settlement appropriate for the future needs and aspirations of citizens, and the best practical arrangements for its implementation.

For more information on the Trust and its Commission, please visit www.2020pst.org.

The views expressed in this report are those of the authors and do not represent the opinion of the Trust or the Commission.
Contents

Authors 5
Foreword 6
Executive Summary 8

1 Introduction 11

2 Why Use Payment-by-Outcome? 17
   What Is It? 17
   Why Is It Used? 19

3 Outcomes 22
   Which Outcomes? 23
   Whose Outcomes? 26
   Level of Outcome 26

4 The Primary Tools: Measures and Standards 29
   Measures, Standards and Incentives 30
   Measures That Work 31
   Designing Standards 35
   How Measures Are Used 39

5 The Primary Tools: Incentives 42
   Adjusting Intensity 42
   Adjusting Diversity 43
   Managing Gaming 47

6 Other Design Tools 50
   Population Segmentation 50
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td><strong>System Design</strong></td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Ownership of the Residual</td>
<td>57</td>
</tr>
<tr>
<td></td>
<td>Building a Learning System</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Designing Procurements</td>
<td>59</td>
</tr>
<tr>
<td>8</td>
<td><strong>Case Study 1: Welfare to Work</strong></td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Population Segmentation</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Designing Incentives</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>The Procurement Process</td>
<td>68</td>
</tr>
<tr>
<td>9</td>
<td><strong>Case Study 2: Offender Management</strong></td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Background</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Payment-by-Outcome Pilots</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>Alternative Service Models</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Selecting Measures</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Controlled Innovation</td>
<td>78</td>
</tr>
<tr>
<td>10</td>
<td><strong>Case Study 3: Long-Term Condition...</strong></td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Background</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Why Pay for Outcomes?</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Tools</td>
<td>83</td>
</tr>
<tr>
<td>11</td>
<td><strong>Other Case Studies</strong></td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>Pharmaceutical Pricing</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>Foster Care</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>Foreign Aid</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Streetscene Management</td>
<td>98</td>
</tr>
<tr>
<td>12</td>
<td><strong>Conclusion</strong></td>
<td>101</td>
</tr>
<tr>
<td></td>
<td><strong>Acknowledgements</strong></td>
<td>103</td>
</tr>
<tr>
<td></td>
<td><strong>Endnotes</strong></td>
<td>105</td>
</tr>
</tbody>
</table>
Authors

This study was undertaken by Gary L. Sturgess and Lauren M. Cumming with case studies by James Dicker, Alexis Sotiropoulos and Nadiya Sultan. The time of the senior authors was a donation to 2020 PST from The Serco Institute, a think tank specialising in the study of public service markets. Two research assistants were employed using donations to 2020 PST by Local Partnerships, Partnerships UK and Serco Civil Government.

Gary has been Executive Director of The Serco Institute since January 2003, driving its research and publication agenda. He is former Cabinet Secretary in the New South Wales state government in Sydney, Australia.

Lauren is a Policy Analyst at The Serco Institute. Previously, Lauren was a Researcher to the Commission on 2020 Public Services. Lauren has a Masters in International Security from Sciences Po Paris and an MSc with Distinction in International Political Economy from the London School of Economics.

James was a Research Assistant at the Serco Institute. Previously, he worked at the Economic and Social Research Council and interned at the Commission on 2020 Public Services and the New Policy Institute. James has an MSc with Distinction from the London School of Economics and was the Adam Smith Prize winner for his BA at the University of Exeter.

Alexis is a Policy Analyst at The Serco Institute, having joined in August 2006. Previously, he worked for the National Offender Management Service and, prior to that, the Department of Trade and Industry. He has a law degree from the University of Cambridge and an MA in legal and political theory from University College London, where he studied at the School of Public Policy.

Nadiya is a graduate intern with Adam Smith International (ASI), working primarily on public financial management within ASI’s government reform practice. Previously, Nadiya was a Research Assistant with the Serco Institute, a position she took up after completing her Master in Public Policy and Management from the London School of Economics with High Merit standing.
Foreword

In the Foreword to this report’s predecessor, *Better Outcomes*, I reflected that ‘Reducing the costs of inputs has value, but if the problem is how better to convert inputs to outcomes, then competition around the costs of the inputs will not solve it.’ Given the fiscal climate in which we find ourselves, that statement is worth reiterating. This report is not about cost cutting and it is not about the benefits of privatisation. Rather, it is dedicated to understanding how government can use payment-by-outcome to extract better results from public services, promote innovation, increase accountability and encourage co-production from service users.

Having produced one report on what we then called outcome commissioning, we felt that there was more work to be done. *Better Outcomes* made the case for payment-by-outcome and explored areas where it could be applied, but because we wanted to keep the report brief and make it accessible to readers unfamiliar with these ideas, discussion of implementation was kept to a minimum. Events have moved on, and policymakers across government are now grappling with the practicalities of how to make payment-by-outcome actually work to reduce unemployment, rehabilitate offenders more effectively, improve clinical outcomes for patients with mental health problems and improve the life chances of the poorest children.

This report, therefore, is very much a toolkit, designed to help commissioners think through the challenges of using payment-by-outcome and how these can be overcome. Through a number of real-world case studies, the authors have identified common problems and solutions that have been implemented or attempted. It does not claim to have all the answers. Rather, it is intended to stimulate further investigation into how payment-by-outcome, with all its obvious benefits and evident difficulties, can be made to work.

The tone of the report is optimistic, but cautiously so, which drives the authors to focus on the practical issues of implementation. While recognising the advantages
of payment-by-outcome, the authors pay serious attention to the criticism that has come from respected thinkers such as Aaron Wildavsky, Allen Schick and Henry Mintzberg. Paying for outcomes is not a panacea and needs to be implemented with due consideration to the goals of the commissioning body and the particularities and context of each service. However, the authors are clear that, done well, it can deliver substantial improvements to public services.

I hope that policymakers, commissioners and public service providers will engage with the report and its authors, so we can continue to learn from experience, share our knowledge and improve our public services.

Lord Geoffrey Filkin
Chair, 2020 Public Services Trust
Executive Summary

Payment-by-outcome is a form of performance management where providers are paid on the basis of outcomes rather than effort. It combines a high-stake form of performance contracting that has come to be known as payment-by-results, with intense focus on the primary outcomes for which government programmes have been introduced.

*Payment by Outcome* is a toolbox. It explores the challenges involved in using this complex form of performance contracting. It seeks to understand the tools that have been employed in the past to cope with these challenges. And it imagines ways in which these instruments might be used differently in the future. It does this by examining the use of payment-by-outcome in welfare to work, offender management and long-term condition management, and drawing on additional insights from several other areas, including pharmaceutical pricing and foster care.

In the public sector, programme objectives are often ambiguous, with primary outcomes surrounded by a variety of contextual goals. Having established the need for early resolution of these ambiguities, *Payment by Outcome* turns to the tools that lie at the heart of this particular form of performance contracting, measures, standards and incentives, arguing that in the design and application of these instruments, commissioners must give full regard to the human dimensions of performance regimes.

Successful performance contracting draws on a wide range of non-financial incentives, so that in designing a system based on payment-by-outcome, commissioners can adjust both the diversity as well as the intensity of incentives. They can also adjust the definition of the service population and the duration of contracts to further improve the alignment of the providers’ and commissioners’ interests.

However, commissioners also have more strategic instruments at their disposal, enabling them to adjust the design of the system within which contracts are made. The most complex payment-by-outcome systems currently in operation – those...
assisting the long-term unemployed into work – have been under development for more than 60 years. While there are obvious benefits to be gained from studying past experience, policymakers must also be patient, recognising that any move to high-stake performance incentives and outcome specification will inevitably be a process of discovery, with initial mistakes and misunderstandings.

It follows that the ideal set of performance incentives and the most effective segmentation of the population cannot be known in advance, but will only be discovered through systemic learning. The report argues that commissioners should deliberately create adaptive systems from the outset, where exploration is encouraged and learning is embraced.

The third high-level conclusion of this report is that payment-by-outcome works best in public services where there are ‘known unknowns’. Where the linkages between inputs and outcomes are well understood and tightly connected, there is little point in specifying outcomes. Commissioners might as well purchase the key inputs or processes that they know will deliver the desired outcomes. On the other hand, where these linkages are so poorly understood that there is very little agreement about the relationship between effort and outcome, it will be virtually impossible to write an outcome-based contract that effectively transfers risk. Under such conditions, providers might just as easily be penalised for failings over which they had no control, or rewarded for successes to which they made only a small contribution.

Payment-by-outcome seems to work best in circumstances where commissioners already have some confidence about the service models that are likely to work, but lack confidence about the capacity of existing delivery chains to deliver significantly better outcomes. Much of the interest in payment-by-outcome seems to relate to certain kinds of innovation: (i) identification of those beneficiaries for whom particular service models will work best; (ii) creation of effective management processes (for example, through joining up fragmented supply chains) enabling services to be tailored to different classes of beneficiary; and (iii) encouragement of much greater co-production on the part of beneficiaries.

The final high-level conclusion from the report is that payment-by-outcome is not always appropriate. Policymakers must be clear about the policy challenge they are grappling with and the nature of the interventions that are most likely to work. If payment-by-outcome is the default option, social problems for which it is not the appropriate solution may be overlooked. When the only available tool is a hammer, there is a danger that every problem looks like a nail.
In some cases, there will be such a long delay between intervention and impact that payment-by-outcome will simply not work. Since the need for long-term investment in intractable social problems is one of the principal reasons why governments become involved, this may be a significant constraint on the use of payment-by-outcome. At other times, the use of a term contract will create a threshold problem at contract termination, with the commissioner and not the provider owning the residual benefits of successful intervention. In such cases, choice-based markets rather than competitively-tendered ones may be more appropriate.

Having been written as a toolkit, much of the report’s value is to be found in the detail, which cannot easily be written up into an executive summary. Its real worth will lie with commissioners who rummage through it and come back to it from time to time for new insights as they grapple with real-world challenges of policy design.
Introduction

Lots of animals, particularly apes, use objects; but what sets us apart from them is that we make tools before we need them, and once we have used them we keep them to use again. This chipped stone from Olduvai Gorge [manufactured around 2 million years ago] is the beginning of the toolbox.

– Neil MacGregor, A History of the World in 100 Objects

This report is the beginning of a toolbox for policymakers charged with commissioning public services from providers who are paid on the basis of outcome rather than effort. It does not seek to solve specific policy problems. It does not offer blueprints for the implementation of payment-by-outcome in particular public services. Rather, by reflecting on the experience of others, it seeks to explore the challenges involved in using this highly complex form of performance management, to understand some of the tools that have been used to overcome these challenges in the past, and to imagine ways in which these instruments might be used differently in the future.

Payment-by-outcome has many names – cash on delivery, payment-by-results, commissioning for outcomes, payment for progress, no cure-no pay – and it is being explored as the possible solution to a wide range of social reforms – finding jobs for the long-term unemployed; raising lifetime earnings of job training participants; breaking the cycle of re-offending; placing children more promptly in stable foster care; rolling out expensive new pharmaceuticals; providing greater stability and accountability in international aid.

The challenge lies in knowing how to use performance targets and financial incentives to focus management on a limited number of high-level policy priorities, promote innovation in service delivery while avoiding some of the distortions associated with traditional performance regimes.
It is regarded as one of the hottest new ideas in public service management today, and yet the principles underlying payment-by-outcome are not new. Over the past 40 or 50 years, performance management with a focus on programme objectives has been attempted in various ways under a variety of different titles – ‘Management by Objectives’, ‘Program Budgeting’, ‘Performance Budgeting’ and ‘Management by Results’.

For the most part, these initiatives have not been a great success, and some of the most highly respected thinkers in the field of management and public administration have warned against them. In 1969, Aaron Wildavsky concluded that: [Program budgeting] resembles nothing so much as a Rube Goldberg apparatus in which the operations performed bear little relation to the output achieved.1 In 1986, W. Edwards Deming wrote: ‘Focus on outcome (management by numbers, MBO, work standards, met specifications, zero defects, appraisal of performance) must be abolished…’ And in 1996, Henry Mintzberg asked: ‘How many times do we have to come back to this one until we give up?’2

And yet, governments around the world continue to explore payment-by-outcome, and in the United Kingdom it has become the subject of intense interest amongst policymakers and practitioners in the fields of welfare to work and offender management, among others. It is not entirely clear what has stimulated this rebirth, but there are a number of possible explanations:

- Performance contracting has been employed by governments across the English-speaking world for several decades, with increasing complexity. Over time, contracting has developed from the specification of inputs and processes, to the incentivisation of outputs and, in some cases, outcomes. In the UK, for example, contracting for refuse collection has developed into the commissioning for maintenance of local streets and the neighbourhood spaces (including litter collection, verge maintenance and graffiti removal), with some contractors paid in part, based on the perceptions of local residents.

Of course, there have been failures in performance contracting, but there have also been notable successes, and policymakers have become sufficiently confident in their ability to commission complex services to experiment with payment-by-outcome. For example, there is a widespread perception that competition and contracting in prison management in the UK and Australia have delivered a comparable level of quality at significantly lower cost, so that
there is now bipartisan support for exploring payment-by-outcome in reducing re-offending.

- In the United States, a number of state and federal initiatives launched in the 1980s made payment-by-results (for outputs if not outcomes) a condition of funding, the most notable examples being job training and foster care. These programmes have been extensively studied, and while they have not been without their difficulties, the findings have in general been encouraging.

  Building on these programmes, governments in Australia, the United Kingdom, France, Germany and the Netherlands, to mention a few, have experimented with payment-by-results contracting in job placement based on the delivery of intermediate outcomes. And while adjustments are still being made to the incentive regimes, policymakers remain optimistic about the progress being made.

- Over the same period, governments in Europe and North America have used performance management to drive service improvements in health and education (although not usually in conjunction with contracting). These initiatives are widely known and controversial, but continue to attract a high level of interest in the policy community.

- Payment-by-outcome has also been employed, with varying degrees of success, in the introduction of expensive new pharmaceuticals. Large corporations have been prepared to accept significant levels of risk in the final stages of research and development, as new drugs are brought onto the market.

- And in the treatment of chronic health conditions, management techniques developed by managed care organisations in the United States in the 1970s have attracted significant international attention. Drawing on the principles of case management and co-production, coordinated care seems to offer valuable insights into how payment-by-outcome might work.

So, despite the pessimistic conclusions of an earlier generation of scholars, payment-by-outcome continues to lie at the cutting edge of public service reform across the industrialised world. It appears that policymakers and service commissioners have had enough success to believe that they can do more with it.

**A Performance Contracting Framework**

Payment-by-outcome is a form of performance management, and there is no reason to believe that it could not be implemented within an individual government department.
or agency using performance bonuses paid to senior managers. However, this report assumes that it is employed in conjunction with performance contracting; that is, it assumes that payment-by-results is implemented using a contractual model, with a public sector commissioner purchasing specified outcomes or outputs from an independent or semi-independent provider from the public, private or voluntary sectors.

This approach was initially adopted as an aid to clear thinking, particularly on issues such as risk transfer, but it is possible that payment-by-outcome may also work better in a contractual environment. Performance budgeting (which usually has a focus on outcomes or high-level outputs), has been explored by a multitude of governments over recent decades, and it has rarely been a success. In part this is because of the inherent tensions of the budgetary process, in part because of the challenges that a focus on outcomes or outputs brings for existing organisational structures, and in part because of the incremental nature of public sector budgeting.³

However, if we look at payment-by-outcome through a contractual lens, the magnitude of these obstacles is significantly reduced. Contracting can be undertaken outside of the budgetary cycle. It is multi-year rather than annual, resulting in lower transaction costs. It can be used selectively, applied only to those programmes presently capable of being managed for results. It can be utilised for all of the funding allocated for a particular function, rather than just the increment of an existing budget; in this way it is possible to stimulate zero-base budgeting each time the function is re-competed. The provider can be held to account financially for the failure to deliver the contracted results. A great deal more risk can be transferred to providers under a contractual framework than is possible through informal ‘service level agreements’ between government agencies. And the negotiation of a contract focuses debate on operative goals rather than abstract high-level outcomes, forcing policymakers to deal with the problem of multiple and inconsistent objectives.

Thus, performance contracting appears to create a much stronger link between resources and results. It may be possible to employ other forms of performance budgeting to accomplish this, but robust examples are difficult to find. For the purposes of this report, it was decided to use the contractual model as a way of framing the issues in the exploration of payment-by-outcome.

Using the Whole Toolkit
This report is not only concerned with performance measures and financial incentives. One of the clearest insights to have emerged from this project is that
policymakers must draw on the full range of instruments in the toolchest. Successful performance contracting has never relied exclusively on targets and incentives. This is true of individual employment contracts used by corporations to motivate their internal workforce as much as it is of public services commissioned from external providers:

Firms use a variety of incentive instruments… Perhaps the most direct is to pay the agent based on measured performance in a given task or set of tasks. But monitoring is imperfect and costly, enabling only a narrow set of activities to be rewarded effectively this way. Asset ownership is often a broader, more powerful instrument. When an agent owns a set of productive assets, she maintains those assets more effectively… A third major incentive is the design of the job: the tasks included in the job description, the activities that are expressly excluded… and the specification of work rules, working hours, and similar policies that restrict the freedom of the worker.4

The following chapters explore this wider range of instruments – the selection of market models and the design of procurements, the use of non-contractual instruments and non-financial incentives within the contractual framework, as well as other elements of task design, such as contract duration and population segmentation.

Methodology

Payment by Outcome builds on an earlier publication by the 2020 Public Services Trust entitled Better Outcomes, which explored the potential gains from payment-by-outcome and where it might be applied, as well as examining some of the challenges involved in its implementation. This report builds on the first, but goes further in seeking to understand how payment-by-outcome models work, investigating in detail the challenges that practitioners have faced and the tools that have been used to overcome them.

The academic literature on performance measurement and management is vast, and the authors have drawn on it to a certain degree, deliberately focusing on those articles that studied real-world cases and paying less attention to those focused exclusively on theory. This is consistent with the general approach of this report, which is to understand how to make payment-by-outcome work in practice.
The principal methodology employed in the research, however, was a case study approach. Three case studies in particular, where payment-by-outcome or other complex performance incentives have been employed in recent decades or are currently being explored, formed the foundation from which insights were extracted and applied to other areas.

a. Job placement and job training – one of the earliest experiments in payment-by-results, and the service where the most advanced examples of paying for outcomes are to be found;

b. Offender management – where the UK government has a firm commitment to introducing payment-by-outcome, where significant design work has already been undertaken by public officials and research institutions, and where several pilots have been launched;

c. Management of long-term health conditions – where there is a history of initiatives involving performance measurement, case management and co-production, and there are insights to be gained from closer analysis.

Several other examples were also briefly reviewed: pharmaceutical pricing, foster care, international aid and streetscene management. All case studies were themselves a combination of desk research, interviews with practitioners involved in the delivery of payment-by-outcome models in the United Kingdom, and conversations with policymakers concerned with the development of new applications of this approach in the UK.

-payment by Outcome is not concerned with answers; rather the cases were studied with the hope of designing a better set of questions. The overriding objective was to understand the challenges involved in developing and implementing payment-by-outcome systems, and to identify the tools that commissioners have employed as they have sought to resolve them. Each case study is different, and comparison has served to highlight some of the reasons why some tools work in some situations and not in others.
2
Why Use Payment-by-Outcome?

Payment-by-outcome will not always be the best solution. There are some public services for which high-intensity performance incentives are not appropriate, for example, where neither commissioners nor providers understand a great deal about the best linkages between inputs, outputs and outcomes. And there are some public services for which the identification and specification of clear and consistent outcomes is extraordinarily difficult.

Since high-intensity payment-by-outcome is not always appropriate, policymakers need to understand the defining characteristics of this particular set of tools, and engage in honest and open debate about why they are being used.

What Is It?
At the time of publication, the term widely used in the United Kingdom for this form of performance management was ‘Payment by Results’. This phrase had come into use with the election of the Liberal-Conservative Coalition, and largely replaced the term ‘Outcome-based Commissioning’, which had been employed by the previous Labour Government. At the same time, policymakers were exploring the use of invest-to-save as a way of shifting significant risk to providers, requiring them to finance a substantial proportion of the up-front costs of intervention, with reimbursement being made later out of identified savings.

Thus, the policy dialogue about performance contracting at the time of publication involved three distinct but related ideas:

a. Payment-by-results, the basic concept, is performance management with teeth, performance budgeting with effective transfer of risk. The term implies that
providers have sufficient control over the resources that impact on outputs or outcomes to be financially accountable if the desired results do not eventuate. It is often assumed that the performance incentives will be high-intensity, with significant financial risk for the provider, however this is not an essential feature of the model, and in some cases may be highly undesirable.

b. **Outcome commissioning** focuses on the specification and incentivisation of the ultimate policy or programme objectives. While it may not be possible to design performance measures that directly drive the delivery of primary outcomes, outcome commissioning seeks to clarify the ultimate ends for which a programme was established, and if necessary, to incentivise the delivery of intermediate outcomes or high-level outputs that serve as surrogates for these primary goals.

c. **Invest-to-save** requires providers to finance the up-front investment necessary for service improvement, with compensation being made out of savings that are directly attributable to that investment. In the case of job placement schemes, providers finance the costs of assisting clients into sustainable jobs, with reimbursement being funded out of reductions in spending on unemployment benefits. Much of the interest in invest-to-save is being driven by current budgetary constraints; however, there is also interest in the concept of ‘service PFIs’, a reference to the Private Finance Initiative, where private providers designed, built, financed and operated public infrastructure such as schools and hospitals, with payment being deferred until the buildings were operational.

There are numerous examples in recent years where these three approaches have been pursued in isolation. For example, ‘Payment by Results’ (PbR) has been employed in the National Health Service to refer to a fixed price tariff introduced for the payment of certain medical procedures in hospitals. These are standardised across the hospital system and are based on national averages. However, this was primarily intended as a system of cost control, without particular regard to ultimate health outcomes or concern with invest-to-save. Under the previous government, Public Service Agreements were grounded in the principal objectives of programmes, but there was no thought that these would be associated with performance incentives. And ‘Invest to Save’ grant funding was used by UK central government for some years (for example, in an attempt to break the cycle of child poverty), but there was no expectation that providers would assume the risk of
delivering demonstrable savings before they were paid. In the UK, these concepts have developed separately from one another.

Over the past year or two, however, politicians and policymakers have begun to bring these concepts together into an integrated framework that would see providers paid, at least in part, based on their success in delivering agreed outcomes. In the two programmes where thinking about this approach is most advanced – welfare to work and offender management – there is also a firm commitment to combining payment-by-outcome with an element of invest-to-save. This is most evident in the Invitation to Tender for the UK’s proposed new Work Programme, released in late December 2010.5

This report is primarily concerned with the first two of these concepts, payment-by-results and outcome commissioning, which explains the use of the term, payment-by-outcome. Invest-to-save is not essential to this kind of performance contracting, and at this stage, it is unclear how extensively this high-stake form of service contracting can be applied. Placing such a large proportion of revenue at risk is certainly not necessary in order to align providers’ interests with those of commissioners, and for the sake of clarity and brevity, consideration of invest-to-save has not been considered at any length in this report.

Why Is It Used?

If payment-by-outcome is an instrument rather than an ideology, then commissioners must understand how it works, the conditions under which it works best, and what they are hoping to accomplish through its implementation. This last question matters since it will rarely be possible to use payment-by-outcome in an undiluted form, and as compromises are made, commissioners must keep firmly in mind what they are hoping to accomplish.

For example, if what commissioners most want is to stimulate greater innovation in the range of service models, exploring alternative linkages between outputs and outcomes, then it may be necessary to limit the amount of financial risk transferred to providers. If, on the other hand, there is a desire to have providers fund much of the investment in service improvement up front, then commissioners may have to be satisfied with service models that are more familiar and thus more easily priced and financed.

Similar compromises will have to be made in the selection of outcomes, the identification of workable measures and the design of incentives – there is no one
model that will work well in every case and commissioners must understand which aspects matter most. This study has concluded that payment-by-outcome is being deployed for a variety of reasons:

- **Transferring performance risk to providers**: There is no point in attempting to contract for outcomes unless commissioners believe that it is possible to transfer performance risk from policymakers to providers, and that there will be sufficient social and economic benefits from so doing. This assumes that there is something yet to be gained, either through the discovery of more effective linkages between inputs and outcomes, or through more effective implementation of what is already known. Payment-by-outcome is being deployed in the realm of ‘known unknowns’, where commissioners believe that there are significant gains to be made, but where traditional managerial approaches to reform are regarded as inadequate.

- **Innovation**: In some cases, government agencies focus on outcome commissioning for the explicit purpose of encouraging much greater innovation in the service models. The effect is to shift responsibility for the linkages between inputs and outputs and outputs and outcomes from policymakers to practitioners. This is most evident in recent discussions of its use in offender management where a great deal remains to be learned about desistance and its practical application. Invest-to-save may have more limited application in such cases, at least in the short term, when commissioners will wish to increase the amount of outcome risk transferred to providers.

- **Devolving authority to the front line**: Even under traditional delivery models, front-line public servants enjoy significant discretion in how policy is interpreted and applied, and where improved service outcomes depend on much greater personalisation of services, the local knowledge possessed by ‘street-level bureaucrats’ acquires even greater value. Payment-by-outcome provides a vehicle for radical devolution to providers, within a framework of enhanced accountability.6

Performance budgeting and performance contracting are accountability reforms that have often been driven by funding agencies, and officials from these agencies have sometimes clashed with policymakers wanting to give local authorities, non-profit organisations and front-line delivery agents greater freedom in operational decision-making. Payment-by-outcome is of interest
in some quarters because it promises to respect the values of autonomy and accountability simultaneously. This is one of the main reasons why it is being discussed in the context of foreign aid and non-profit funding.\textsuperscript{7}

- **Secured funding:** Because it implies such a strong accountability framework, payment-by-outcome may provide funders with greater confidence in the capacity of providers to deliver measurable outcomes, and thus attract additional funding in the short term or a firmer commitment to financial support over the longer term. Security of this kind will be of interest to commissioners seeking to make strategic reforms. This appears to be one of the benefits of invest-to-save, where Treasury may be encouraged to release additional funding in future years once the evidence of cashable savings is already clear.

- **Joining up funding streams:** In some cases, payment-by-outcome might be used as a way of integrating funding streams. Sharper focus on primary outcomes enables commissioners to ask challenging questions about the potential benefits of joint funding of shared objectives, and agencies with a strong case for joining-up on their terms can also be expected to support payment-by-outcome for this reason. Invest-to-save may also facilitate integration, since agencies will not need to draw on existing budgets in order to fund such an initiative.

- **Integrated delivery:** There are also those who argue that outcome commissioning will facilitate the joining up of service provision. This may be assisted by the joining up of funding streams, but even where that cannot be accomplished, the transfer of outcome risk to providers should create stronger incentives to bring together those services that necessarily must be interconnected to ensure that key outcomes are delivered.

- **Overcoming distortions in performance management:** Finally, there is some support for payment-by-outcome among providers and commissioners who recognise the benefits of payment-by-results, but also acknowledge some of the difficulties in designing effective performance regimes. In this case, it is the focus on outcomes that offers the greatest scope for improvement.
3 Outcomes

The challenges involved in designing and operating effective performance regimes are well understood and can be said to lie in ‘the folly of rewarding A, while hoping for B’. Payment-by-outcome seems to avoid this folly by defining what is required (and thus, what will be rewarded) in terms of social impact, rather than the inputs or outputs that contribute to those ends.

However, the use of outcomes as performance measures is also problematic, for a variety of different reasons. As James Q. Wilson noted some years ago:

Outcomes – results – may be hard to observe because the organization lacks a method for gathering information about the consequences of its actions (for example, a suicide-prevention agency may actually prevent suicides but it has no way of counting the number of potential suicides that did not occur); because the operator lacks a proven means to produce an outcome (for example, prison psychologists do not know how to rehabilitate criminals); because the outcome results from an unknown combination of operator behavior and other factors (for example, a child’s score on a test reflects some mix of pupil intelligence, parental influence, and teacher skill); or because the outcome appears after a long delay (for example, the penalty imposed on a criminal may lead to a reduction – or even an increase – in the offender’s behavior five years later).

After studying the measurement of clinical outcomes in the health sector, Professor Peter Smith of Imperial College London arrived at similar conclusions, arguing that outcome measurement works best where: the nature of the outcome is relatively uncontested; it can be captured relatively easily in operational performance
measures; an indicator of the outcome can be secured reasonably soon after the intervention; the outcome is readily attributable to clinical performance rather than external factors; and there is need for considerable clinical judgement as to the most appropriate intervention to offer.\textsuperscript{10}

Thus outcomes are not always suitable for use as performance indicators and it is often necessary to use outputs as proxy measures to give the commissioner some confidence that the primary objective is being delivered. In cases where measuring outputs is too complex, it may be necessary to specify processes or inputs.

The challenge of translating primary goals into workable measures brings back the risk of creating byzantine rules and perverse incentives that fail to deliver what policymakers ultimately want – hoping for B whilst incentivising A. Thus, while outcome commissioning may ameliorate the challenges involved in designing effective performance regimes, by no means does it eliminate them entirely. It follows that commissioners must be absolutely clear about what they want to achieve from a particular programme.

**Which Outcomes?**

Identifying the primary outcome of a particular programme or agency can itself be problematic. Wilson argued that public agencies often have contextual goals, ‘descriptions of desired states of affairs other than the one the agency was brought into being to create’.\textsuperscript{11} In part, this is because policymakers need to garner support from broad constituencies, and this helps to explain the widespread existence of outcomes that are stated in ambiguous and inconsistent terms.

This is sometimes evident in the establishment of new programmes, even where they rely on payment-by-results. In the United States, the Job Training Partnership Act, a performance-based scheme legislated by the US federal government in 1982, focused training on ‘those who can benefit from, and are most in need of, such opportunities’. Service providers were simultaneously directed to focus on equity (those most in need of assistance) and efficiency (those who would most readily benefit from assistance), and since there was no reason to believe that these would be the same individuals, conflict was introduced from the outset as to how resources should be prioritised.\textsuperscript{12}

With long-established programmes, contextual goals have often accreted over time through incremental policy adjustments. One of the distinct advantages of
payment-by-outcome is that it introduces greater discipline into thinking about objectives, since providers will be reluctant to accept the financial risk of delivering results that have been expressed in ambiguous language or remain the subject of internal conflict. Where a department or agency has been delivering services over many years, this complexity may not be obvious, and it is unsurprising that a move to performance contracting sometimes results in tension, as commissioners stumble across implicit outcomes for which there are entrenched but previously unrecognised constituencies.

Where services are commissioned across organisational boundaries, then different commissioners and/or providers will often have conflicting missions and cultures. However, ambiguity or inconsistency may also arise from differences within an agency. Fifty years ago, the North American sociologist, Charles Perrow noted that:

Official goals are purposely vague and general and do not indicate two major factors which influence organizational behavior: the host of decisions that must be made among alternative ways of achieving official goals and the priority of multiple goals, and the many unofficial goals pursued by groups within the organization.13

Perrow described these unofficial objectives as ‘operative goals’, and argued that they could only be discerned through careful observation. One of the risks involved in assuming that primary outcomes are unambiguous and that delivery is simply a matter of identifying the appropriate measures and incentives is the temptation to conclude that the only real obstacle to success lies in manoeuvring by middle management for status, resources and power.14

Moreover, attempts to use payment-by-outcome as a vehicle for joining up fragmented supply chains may reinforce the temptation to specify outcomes at a highly abstract level, and encourage commissioners to overlook the brute reality that, in order to be effective, payment-by-outcome must exclude as well as include. As Pressman and Wildavsky observed:

If we relax the assumption that a common purpose is involved... and admit the possibility (indeed, the likelihood) of conflict over goals, then coordination becomes another term for coercion.15
One of the defining features of performance management systems is that they demand a focus on specific results that are regarded by policymakers as having a higher priority, and by definition, this must involve downgrading or excluding other objectives. This has been particularly evident, for example, in the controversy over education reform, where performance measures have directed schools and teachers to concentrate on literacy and numeracy at the expense of other important aspects of learning such as creativity and higher-order problem-solving, which are not as easily measured. This may be what policymakers intended, but if they are to avoid unintended consequences, commissioners must be explicit about the objectives that will be relegated to secondary status, or what compensating measures will be deployed.

One of the core responsibilities of commissioners must be to identify where different individual and organisational missions and objectives are not aligned, and find ways of managing the associated tensions.

That we focus our attention on a particular goal, singling it out as our objective, does not mean there are not others within which we must also operate or, at least, find ways to relax or overcome. Knowing only the avowed programmatic objective without being aware of other constraints is insufficient for predicting or controlling outcomes.¹⁶

Where it is not possible to reduce performance measures to a relatively small number of clear and consistent objectives, it will be difficult to shift the risk of delivery down to operational managers, whether they are employed directly by government or indirectly under contract. Wilson concluded that:

… the more contextual goals and constraints that must be served, the more discretionary authority in an agency is pushed upward to the top... The greater the number and complexity of those goals, the riskier it is to give authority to operators.¹⁷

Thus, if the responsibility for delivering programme objectives is to be devolved to front-line service managers (which appears to be one of the principal benefits of using payment-by-outcome), then it is essential that commissioners acknowledge the complexity of organisational or programme outcomes, and seek to reduce them to a set of unambiguous deliverables.
The pursuit of clarity does not mean that payment-by-outcome models cannot also incorporate complexity and flexibility. However, it is unlikely that high-stake incentive schemes, where significant financial risk is shifted to providers, will be suited to such models. In public service contracting, contextual goals are often addressed through non-financial performance incentives, including corporate reputation, professional norms and governance arrangements, and these tend to result in a less highly leveraged financial incentive regime.

Whose Outcomes?
Another source of ambiguity is created when multiple layers of commissioners are involved in the administration of a performance regime. This is most evident in federal systems, where the responsibility for delivery is sometimes cascaded down from a national government, which supplies funding and policy oversight, to subordinate provincial and local departments and agencies. Each may have priorities that differ from those of the ultimate commissioner, resulting in a proliferation of decision-points, with much greater scope for misunderstanding and misalignment of interests.18

While this may be less of a problem in unitary systems such as the United Kingdom, there are still sufficient layers of administration in, say, the National Heath Service, for ambiguity of this kind to emerge. And the deliberate use of payment-by-outcome as a way of joining up public services will create new sources of uncertainty around the ownership of outcomes.

If payment-by-outcome is to be effective, providers must know whose outcomes are being commissioned. One way of resolving this ambiguity lies in identifying a lead agency. But where multiple stakeholders continue to play a role in the allocation of priorities, providers must have a way of discovering well in advance how multiple objectives will be prioritised, and thus how measures will be interpreted and financial rewards and penalties allocated.

This implies that policy adjustments will be punctuated, with change being made periodically, rather than evolving organically over time. While this has some obvious downsides for policymakers, it has major advantages for providers, and makes it possible to transfer performance risk successfully.

Level of Outcome
Outcomes must be sufficiently abstract to allow for innovation in service delivery, including the integration of previously ‘silod’ services, yet specific enough to enable
independent observation and objective measurement. They must be sufficiently ambitious to encourage innovation, but realistic enough that providers will be willing to assume the risk of delivering the outcomes.\textsuperscript{19}

In an ideal world it would be possible to contract for primary outcomes, with success or failure ascertained using associated performance measures and appropriate financial incentives. In the real world, this will hardly ever be possible. In some cases, primary outcomes may simply be unmeasurable, or they may capture so many extraneous variables that it is impossible to attribute improved outcomes to the interventions of individual providers. In other cases, outcome improvements may only be observable after such a long time period, that a performance contract may be unsuitable because of high monitoring costs and the difficulty of attribution.

- **Measurability** is the principal reason why the primary objective of reduced re-offending is useless as a performance measure. In this case, it is difficult to obtain reliable statistics on the number of offences that are being committed, and impossible to measure the rate of re-offending, since that would require the attribution of known crimes to known offenders. Here it is difficult to imagine what kind of system might generate the required information, since offenders have a strong interest in concealing their own rate of re-offending. In other cases, the methodology might be understood, but the costs of collecting the data might be prohibitive.

- **Duration:** The primary outcome in many welfare to work programmes is sustainable jobs for the long-term unemployed. In the UK, the performance measures used in existing schemes are heavily based on retained employment at 13 and 26 weeks, which is not necessarily a strong indicator of sustained participation over the longer term. It is for this reason that the Department for Work and Pensions has proposed that under the Work Programme, performance periods be significantly extended. By way of example, for claimants of the Job Seekers Allowance (JSA) aged 25 and over, providers will be paid around one quarter of the potential total only if clients are kept in employment for six months, with another two-thirds, paid in four-weekly instalments, for a further period of up to twelve months. For some of those claiming the Employment and Support Allowance (ESA), the maximum period of sustainment will be two years. These measures have been published in the Invitation to Tender issued in December
2010, and it has yet to be established whether a payment-by-outcome and invest-to-save scheme based on such long performance periods is bankable.²⁰

- **Attribution:** Ideally, outcomes should also be set at a level that minimises the contribution of external variables. The contribution of one particular prison or probation manager to lower reconviction rates within a defined population may be small. Commissioners may compensate for this by reducing the intensity of the targets and/or the associated financial incentives, but the challenges of attribution might also be overcome by reducing the level at which performance is measured, from primary or intermediate outcomes, to high-level outputs. Providers will have much greater control over outputs such as successful completion of a drug rehabilitation course, or placement and sustainment in employment.

Regardless of whether commissioners choose to measure and reward performance at the level of primary or intermediate outcomes, high- or low-level outputs, processes or inputs, under outcome-based commissioning, they should commence the design process with a close study of the ultimate objectives for which the programme in question has been established, and ensure that the measures selected are aligned as closely as possible with these goals.
The Primary Tools: Measures and Standards

It has been argued that the dominant philosophy underlying performance incentive models such as payment-by-outcome is the notion of ‘managerial cybernetics’:

In its crudest form this envisages the managerial process as follows. Organizational objectives are identified. Performance indicators are developed to reflect these objectives. Targets are set in terms of the performance indicators. Management then chooses action and effort to achieve the targets. Progress towards targets is monitored using (performance indicators), and – if there is a divergence from targets – new targets are set and appropriate remedial action is taken.21

This passage was written in the early 1990s, drawing on research from the 1960s, and yet it captures almost in its entirety the underlying structure of payment-by-outcome – identification of outcomes, choice of measures to reflect those outcomes, selection of standards and associated performance incentives. The principal flaw with the cybernetic model, as the author of this passage recognised, is that organisations are made up of complex networks of human beings, all able to (and inclined to) exercise their own will.

Performance management has contributed a great deal to organisational efficiency in the public and private sectors over recent decades, but it has been much less effective where it has failed to take into account this human dimension. This report has been written on the assumption that commissioners understand this unavoidable reality, and build it into the design of their performance regimes.
It is not a matter of designing the perfect cybernetic system, and then correcting for gaming once the people are added, but recognising from the outset that the design of measures, standards and incentives must acknowledge the existence of human will. This does not mean that commissioners must anticipate all of these human reactions in advance (although they should certainly try), but rather that they should build adaptive systems that are capable of learning from these responses and being reinvented.

Moreover, in taking the human dimension into account, commissioners will discover that they have available a much wider range of tools than might otherwise be suggested by the cybernetic model, including reputational incentives, organisational culture and professional ethos. These aspects are addressed in a later chapter, but first it is necessary to focus on the primary instruments of payment-by-outcome.

**Measures, Standards and Incentives**

Measures, standards and incentives are the primary, though by no means the only, tools of a performance measurement regime, and it is with these particular instruments that this chapter is concerned.

The selection of measures will be closely informed by decisions about the level at which outcomes or outputs are to be commissioned. Thus, in the case of offender management where measurement of the primary outcome is not possible, commissioners must decide which intermediate outcome (reconviction or re-imprisonment) and/or which high-level outputs (education, employment, drug rehabilitation and so on) will be assessed. Standards are the targets or levels of acceptable performance that providers are expected to achieve before they are rewarded or sanctioned, and these may be expected to change more frequently than the measurement regime.

Incentives are sometimes described as the rewards or penalties applied for meeting or failing to meet these standards, but this narrows the range of potential motivations to monetary ones. On the principle that ‘what gets measured gets done’, it is probably impossible to conceive of a measurement regime that does not generate behavioural incentives of some kind, but commissioners have a wide range of choices about how measures will be used and thus what incentives will be created. A provider’s comparative performance may be published amongst its peers, or league tables may be broadcast to the public at large, causing considerable
harm to its reputation. By comparison, a small financial penalty may not serve as a particularly powerful incentive, but if a substantial part of the provider’s revenue is at risk based on performance, then financial incentives will operate with much higher intensity than reputational ones.

However, measures and incentives are only some of the tools to be found in the commissioner’s toolkit, and how they are employed and with what intensity will inevitably be influenced by the way in which other instruments are used. Contract duration and population segmentation are two such factors, but commissioners are also able to influence the behaviour of providers through the ways in which systems as a whole are designed.

Creating a coherent payment-by-outcome regime demands that commissioners understand how measures and incentives operate, how they are related to the primary outcomes, how they work in conjunction with other tools, how providers may respond when confronted with the full range of incentives, and how to make changes in the face of unintended consequences.

Since it is unfair to expect that commissioners can have a full appreciation of all these dimensions from the outset, they should consciously design an adaptive system, so that lessons are learned, alternative tools are employed, and adjustments are made over time in order that the performance regime delivers the outcomes for which the programme was established.

Measures That Work
One of the greatest potential benefits of payment-by-outcome is that it allows commissioners to specify, measure and reward the ultimate objective, thereby avoiding the distortions associated with the specification of multiple tasks. However, commissioners cannot avoid the onerous responsibility they bear for designing measures that work. In part, this is because primary outcomes are not often directly measurable, but even where they are, commissioners must have a mature understanding of the linkages between effort and outcome to be confident that the measures are appropriate.

Selection of Measures
The story has been told of the perverse incentives created by an incentive contract signed by Ken O’Brien, one of the most highly-ranked quarterbacks in North American football:
Early in his career, he had a tendency to throw interceptions. As a result, he received a contract that penalized him every time he threw a ball to a member of the opposition. However, while it was the case that he subsequently threw fewer interceptions, this was largely because he refused to throw the ball, even in cases where he should have done so. As Joe Namath put it, ‘I see him hold onto the ball more than he should… I don’t like incentive contracts that pertain to numbers.’

As a former quarterback, Namath instinctively recognised that a performance regime that focused on only one of the dimensions of this complex role would inevitably result in perverse incentives. Where the delivery of the desired outcome involves multiple tasks or where a single task has several dimensions, but measurement and compensation are based on a subset thereof, providers will reallocate resources to the sub-set for which they are compensated, rather than delivering the full range of desired activities.

In general, where there are multiple tasks, incentive pay serves not only to allocate risks and to motivate hard work, it also serves to direct the allocation of the agents’ attention among their various duties.

The complexity and multi-dimensionality of public services is less problematic if commissioners measure several aspects of performance instead of attempting to capture the outcome in a single measure. Of course, using too many measures can also create difficulties since it increases the burden of data collection and reporting and could ultimately constrain providers’ ability to innovate and personalise services.

Even where the full range of tasks is incentivised, some may be more observable or easier to measure, and there will be a tendency for these to be prioritised. James Q. Wilson argued that there is a kind of Gresham’s Law at work in many government programmes: ‘Work that produces measurable outcomes tends to drive out work that produces unmeasurable outcomes.’

**Understanding the Service Model**

Where the linkages between outputs and outcomes are well established and widely-agreed, it is easier to design an effective system of performance measurement. This is more likely to be the case in the pharmaceuticals industry where the causal
relationships between taking a particular drug and a given outcome must be well-researched before the drug can be approved. Streetscene management provides an example of a relatively straightforward service in which the correlation between outputs, such as the prompt collection of litter and the timely mowing of verges, and the outcome of resident satisfaction, is relatively simple and easily understood.

However, there are many public services in which outputs are not closely correlated with outcomes, due to such complexities as human agency, the willingness of clients to co-produce, and external interventions such as recessions or pandemics, and in these cases, Schick may be correct when he argues that:

There is no inherent causal link between [outputs and outcomes]. Some outcomes may derive from specified governmental outputs, many do not. Alternatively, producing the right outputs does not ensure that the desired outcomes will materialise.27

The ultimate objective of human happiness is no doubt one example of such an outcome. But the ambition of significantly reducing re-offending may well be another. The leading US criminologist, Joan Petersilia, pointed to this kind of complexity in offender management when she declared: ‘There is nothing in our history of over 100 years of reform that says that we know how to reduce recidivism by more than 15 or 20 percent. And to achieve those rather modest outcomes, you have to get everything right.’28

This makes the selection of good measures inherently difficult. In some cases, this problem can be overcome by specifying intermediate outcomes; for example, in measuring the success of providers in reducing re-offending, commissioners might monitor the reconviction rate. This would be more closely tied to reductions in re-offending than, say, employment outcomes, and would save commissioners the difficult task of attempting to specify all the outputs that may contribute to desistance.

**Measuring Impact**

While policymakers must track progress towards the ultimate outcomes, what commissioners most need to understand when making decisions as to whether or not to reward providers, is the value providers have added or the impact they have made on the achievement of those ends. The difference between outcome and
impact can be explained by external factors, such as the robustness of the general economy or the presence of a pandemic, or the ability of clients to solve their own problems without the assistance of state-financed service providers.

Where the commissioner has not specified the measures and standards so that it is the providers’ impact that is being rewarded, there is scope for intentional or unintentional cream-skimming. Thus, if providers are paid a fee for each person who does not suffer an emergency admission to hospital in a given year, and they have some discretion over which clients they serve, providers will be inclined to select patients with a low risk of admission.

However, if the payment is adjusted so that providers serving high-risk patients are paid more than those focusing on those with a lower risk profile, this perverse incentive will be removed. These kinds of adjustments are usually made based on regression analyses using predictive data, although it is possible to make retrospective adjustments based on actual data, with some or all of the payments delayed until the end of the measurement period.

Under the American Job Training Partnership Act, providers had discretion over which clients they enrolled because inadequate funding meant they were only able to serve less than 5% of the eligible population. In that case, researchers found only modest evidence of creaming. In part, this was because payments were adjusted using a statistical model that took into account the predicted client profile (based on data from previous years) as well as regional economic conditions. Moreover, at one site studied, the ‘social worker mentality’ of front-line staff was so strong that it appears to have trumped any incentives to cream.

Yardstick competition can also be used to measure the impact of providers. If two or more providers operate under the same conditions and serve the same types of clients, the results of their interventions can be compared. The act of comparison eliminates the impact of extraneous variables, so that if one provider delivers fewer outcomes than another, it is likely that they are underperforming. If outcomes are similar but fewer than expected, then in the absence of collusion, this would amount to prima facie evidence that some external factor has hampered performance.

Randomised control trials are the most scientifically rigorous way of measuring value-added, but for ethical and practical reasons they are seldom used in human services.
Individual or Population Outcomes

Individual outcomes are obviously important since they are indicative of concrete, positive change in a person’s life; however, measuring the outcomes of a whole cohort of individuals is usually the better design choice. For example, commissioners have the option of measuring and rewarding providers based on each individual that has entered employment, or measuring the percentage of the assigned population that providers have got into work.

Measuring population outcomes is a precondition for identifying and isolating the impact of extraneous variables, and thus increasing confidence that providers are being paid for the value they have added. The disadvantage of this form of measurement is that it requires monitoring of an entire cohort over time, with providers receiving outcome payments only at the end of that period.

However, in Flexible New Deal, a programme for the long-term unemployed, commissioners were able to rely on the measurement of individual outcomes. Because of the long period of unemployment of these jobseekers, any outcome that occurred while jobseekers were on the Flexible New Deal could be assumed with some confidences to be a result of the programme.

Taking Other Tools into Account

The design of performance measures does not occur in a vacuum. The measures selected will have an impact not only on the incentive regime, but also on other aspects of the entire payment-by-outcome system. In particular, population segmentation should be kept in mind when designing the measurement regime.

Measures which are appropriate for one sub-group may not function well for another. For example, for patients with long-term conditions who are at high risk of hospitalisation, measuring the number of emergency hospital admissions may well be appropriate. For lower-risk patients, a more informative measure may be the rate of healthcare utilisation, including GP, specialist and hospital visits. (The subject of population segmentation is further discussed in section 6.1.)

Designing Standards

Standards establish the level of required achievement, and it is not unusual for multiple standards to be used for a single programme. In the UK Employment Zones, for example, providers had to ensure jobseekers retained work for 13 weeks in order to be eligible for outcome payments. In addition, they were also required to
obtain work for a certain percentage of jobseekers referred to them in order to be eligible for a bonus award (or prize as these awards are sometimes known).

Ideally, standard-setting should be based on a solid understanding of the status quo so that targets are realistic but sufficiently challenging to stimulate innovation on the part of providers. There is a risk that competitive tenders launched without any real understanding of the baseline will result in a ‘winner’s curse’, with the successful competitors bidding a price and/or level of performance risk that is unattainable.

This risk is greater where providers compete in setting the standards. In the UK’s Pathways to Work, for example, the commissioner established the standard for clients remaining in work at 26 weeks, while providers competed to set their own monthly targets for numbers of job and sustained job outcomes. As noted in Chapter 8, Pathways to Work resulted in a winner’s curse, with contracts being signed at levels of performance that were unachievable.

This is not to say that providers cannot play a role in standard-setting, and this is one issue that could be addressed in consultation with the industry and through competitive dialogue. Inevitably there will be tension, since providers would prefer standards to be ‘realistic’, while commissioners would like them to be stretching so as to encourage innovation.

**Types of Standards**

The type of performance standard selected can have a large impact on the incentives that providers face. There are at least three different kinds. *Thresholds* are strict standards in which providers receive no reward until a certain level of achievement has been reached. For example, in one streetscene management contract in the UK, the provider did not receive any of the outcome payment unless 60% of the area’s residents were satisfied with the services they received (although the outcome payment was a small proportion of total revenue).

Thresholds have the advantage of setting high standards and giving providers a clear focus, but can also create large distortions. As Propper and Wilson warn, ‘target indicators introduce an arbitrary dichotomy into continuous data and will therefore focus agents’ attention on the borderline’. In the education sector, thresholds have been blamed for focusing teachers’ attention on pupils who, without extra help, may have just missed the target, at the expense of pupils who would have had greater difficulty in meeting it, and thus were the kind of student the policy was probably meant to assist.
With *distance-travelled* standards, the aim is to reward incremental improvement, rather than setting the bar at a particular height. Proposals for payment-by-outcome models in offender management often advocate the use of distance-travelled standards, so providers would be rewarded for improvements on the current re-offending rate, rather than for lowering re-offending to a certain level. This approach is adopted in part because desistance is not often binary, and offenders typically reduce their rate of re-offending over time.

Distance-travelled payments often relate to the time period over which outcomes are sustained. The performance regime for the UK’s proposed new Work Programme is based in part on monthly payments for a defined period of time over which employment is sustained. These standards are especially useful where the status quo is not well understood. The disadvantage of using such standards is that they may not be sufficiently challenging to incentivise providers to innovate and achieve far better outcomes.

*Milestones* can be used with intermediate outcomes, in which case they signal progress against a distance-travelled measure, or with outputs, where they represent stepping stones on the road to intermediate or primary outcomes. They measure progress that commissioners value in its own right and choose to reward. This may be because the milestones signal the achievement of a second outcome (such as lower costs), or because they reduce performance risk for providers or assist them in dealing with cash flow problems, or because they provide early evidence of part-performance and thus help to overcome problems with creaming.

In the UK, welfare to work providers receive payments when jobseekers are placed in a job and then when they retain work for 13 weeks, even though the primary focus is on ensuring that clients remain in employment for 26 weeks. These milestones are considered valuable enough to be worth rewarding, perhaps because they provide evidence of cost savings, although they have the additional advantage of addressing cash flow requirements on the part of providers (particularly important for small to medium-sized enterprises and not-for-profit providers). Milestone payments can also serve to motivate providers to work with disadvantaged clients who may not meet the final standard but will nevertheless achieve some of the benchmarks.

The problem with milestones lies in selection. In some cases, payment-by-outcome may be used where commissioners are unsure of the linkages between outputs and outcomes, so it is difficult to know how to reward progress. In other
instances, personalisation is required and rewarding the achievement of particular outputs can detract from this.

Of course, these various types of performance standards can be combined, and often are. In the payment-by-outcome pilot for Peterborough Prison, a blend of threshold and distance-travelled measures is being used. The investors only begin to make a return when re-offending has decreased by 7.5% on the current rate; after this point they make increasing returns, up to a maximum reduction in re-offending of 13%.

In all cases, rewarding a provider’s contribution to achieving the standard will create fewer distortions than simply rewarding the provider because the standard was reached. The challenge lies in knowing how to measure providers’ impact. Consider a situation in which providers have discretion over which jobseekers they accept onto their programmes. If providers are simply rewarded for each jobseeker obtaining work, they may attempt to screen potential clients based on their probability of finding jobs, leading to inequality of access. If, however, providers’ rewards for job placement are adjusted based on the characteristics of the clients they serve (clients’ distance from the labour market), they will not be incentivised to restrict programme access.

**Duration of Performance Period**

The length of time over which providers must sustain their performance in order to qualify for payment is another variable that commissioners must take into account when setting performance standards. In some cases, the outcomes public services aim to achieve are only deliverable over the very long-term. Indeed, there are many services that are commissioned by government rather than private individuals or families for the very reason that benefits will only be realised over the course of a lifetime.

Payment-by-outcome does not work well with policies where there is a long delay between action and outcome. Early intervention schemes are a classic example, where it is hoped that support services provided to children and parents in the early years of life will improve health and education outcomes in later life. Time lags may also pose a problem for the use of payment-by-outcome with certain categories of criminal offender, although this remains the subject of ongoing research and debate.

The reasons why a long delay between intervention and measure matter are several. First, extraneous variables will become more problematic as the
performance period is extended; it will become more difficult to determine the impact of the initial intervention on the status of the individual, as indicated by the outcome measure. Second, outcome payments cannot be withheld for very long periods if commissioners wish to sustain a diverse and commercially viable support market. Finally, over extended periods, monitoring costs will outweigh the benefit of ensuring the outcome has truly been achieved. This means that where the results of interventions are not visible for many years, as in early intervention services, paying for outcomes will not be suitable.

Where payment-by-outcome is implemented, commissioners must strike a balance between a long monitoring period, which may increase the certainty that the primary outcome has actually been delivered, and the practicalities of extraneous variables, provider cash flow and monitoring costs. To do this well requires a solid understanding of the intermediate and primary (or in this case, short- and long-term) outcomes.

How Measures Are Used

Measures and standards are capable of being used in a number of different ways, and it is not necessary that they should always be expressed as mathematical formulae. For example, in thinking about performance budgeting, the Canadian academic, Allen Schick, has drawn a distinction between their use as analytic tools and as decision rules:

… analytic tools empower budget makers, whereas decision rules constrain them. The former allow full scope for judgment and subjectivity, the latter made budgeting less judgmental and more objective.32

Another way of making this distinction is to contrast the use of measures for weighing performance as opposed to just counting it. The difference between the two lies in the interposition of human discretion, and relatively crude performance measures can be highly effective as long as they are balanced by human judgement. Freelance journalists are typically paid by the word, but this does not operate as a perverse incentive, since payment is based on the words published, not on the number submitted. The editor has complete discretion over which words are finally used, so that what appears to be a crude quantitative measure operates in practice as a highly-effective form of quality control. Thus, measures and standards can
operate either as an input to professional decision-making by contract monitors, or as formulae with which monitors are invariably required to comply.

Measures also work differently when they are used as decision rules, with several financial or reputational penalties attached, leading some observers to conclude that ‘when a measure becomes a target, it ceases to be a good measure’.33 Pascal and Marschke give the following example from an automotive repair chain in North America.

Before Sears implemented its incentive scheme, Auto Center profits may have been positively correlated with the number of repair jobs. This statistical relationship may have prompted Sears officials to compensate Auto Centers on the basis of the number of repairs completed. Once Sears began paying managers bonuses for meeting service quotas, however, those service quotas became the managers’ objective. It was not long before the managers had found easy ways to boost sales volumes that did not also result in higher store profits. By charging customers for un-needed and unperformed repairs, store staff uncoupled the performance measure from the store’s long-term profits. Their response to the incentives drove up the value of the performance measures while driving down profits. Thus, repairs and long-term profits would not be positively correlated after Sears based pay on the number of repairs performed.34

While the publication of numerical performance standards can create high-stake incentives for providers, in general, reputational incentives will provide commissioners with much greater flexibility. Australia’s Job Network ‘star rating’ system is not used to determine how much providers are paid, but to decide whether to extend or automatically renew existing contracts. In this way, star ratings operate as a particularly powerful reputational incentive. While they are quantitative in nature, they do give attention to certain qualitative dimensions, such as providers’ performance in serving the most disadvantaged.

It is also possible for commissioners to construct complex assessment frameworks with relatively few numerical measures. For example, HM Inspectorate of Prisons for England and Wales employs robust and long-established assessment criteria when it inspects prisons and young offender institutions, but these are subjective in nature, and laid down in a 248-page guidance document entitled
‘Expectations’, now in its third edition. While the inspections are based on a large number of assessments about processes or low-level outputs, the Chief Inspector’s published reports provide an overall judgment about the quality of the prison, based on a simple four-part ranking.
In selecting the performance incentives to be used in conjunction with the measures and standards, commissioners are able to adjust their intensity – the level of risk assumed by providers – and to select among a range of different incentive mechanisms that operate in somewhat different ways. Contrary to what is often assumed, commissioners have a great deal of choice in the design of the incentive regime and they are not forced to rely exclusively on financial rewards and penalties.

### Adjusting Intensity

Any distortions in the design of a measurement regime will be magnified if the associated incentives are high-stake, with the potential for large gains or losses on the part of providers. Invest-to-save schemes are high-stake, since providers are expected to finance a substantial proportion of the investment in improved outcomes (and thus reduced costs) before any payment is received. However, any system of performance incentives can be constructed so that it is high-intensity, if a sufficient proportion of revenue is at risk, or there is the potential for a large enough impact on reputation.

Research into the impact of high-stake incentives on teaching in the US has found (unsurprisingly) that:

> … as stakes increase, so does the influence of the test... teachers in high-stakes situations... reported feeling more pressure to have their students do well on the test, to align their instruction with the test, to engage in more test preparation, and so forth.\(^{35}\)

Given that the objective of state-mandated testing was to focus school districts, schools and teachers on specific outcomes, this might indicate that the policy has
been a success. However, increasing the severity of the consequences also creates stronger incentives to ‘teach to the test’:

Teachers in states with high-stake tests are much more apt than their counterparts in states with lower-stake tests to engage in test preparation earlier in the school year; spend more time on such initiatives; target special groups of students for more intense preparation; use materials that closely resemble the test; use commercially or state-developed test-specific preparation materials; use released items from the state test; and try to motivate their students to do well on the state test.\(^{36}\)

Using previous tests to drill students helps to improve test results, but research suggests that it is unlikely to contribute to meaningful learning. This comes as no surprise, but serves as a reminder that as the stakes are increased through higher financial rewards and sanctions, any distortions in the measurement regime are magnified and the incentive to game is strengthened.

High-stake incentive regimes are problematic where the nature of the service performed is complex and quality is difficult to measure. Thus, with the planting of seedlings in forestry plantations, firms in British Columbia generally pay their workers piece rates rather than salaries since these are associated with significantly higher productivity. However, where planting conditions are difficult, so that workers are required to exercise more care and discretion to improve the seedling’s chances of survival, salaries seem to be used more often.\(^{37}\)

As tasks become more complex, greater collaboration is required among different providers, outcomes become more difficult to monitor and greater judgement is expected of providers, commissioners tend to reduce the intensity of their performance regimes and sometimes withdraw from the use of performance incentives altogether.\(^{38}\)

Adjusting Diversity

Incentives must be considered in their totality: ‘the range of instruments that can be used to control an agent’s performance in one activity is much wider than just deciding how to pay for performance’.\(^{39}\) Much of the academic literature seems to assume that financial penalties are the only tools in the commissioner’s toolkit. Were this assumption to be built into the design of a payment-by-outcome system,
commissioners would have little alternative but to ratchet up the intensity of incentives whenever they wished to focus providers more closely on desired outcomes.

In fact, successful performance contracting usually relies on a blend of incentives to align the interests of providers and commissioners. Thus, in a mature payment-by-outcome system, we should expect to find a reliance on financial rewards as well as financial penalties, regulation and certification, individual and corporate reputation, professional norms and organisational culture, and structure and governance.

- **Financial rewards:** Psychologists maintain that fear is a more powerful source of motivation than hope, and this, combined with the fact that in political debate it is easier to be seen as punishing failure than rewarding success, results in a systematic bias in favour of using penalties. However, it is difficult for providers to encourage their employees to strive for excellence when success is never acknowledged, and the only form of public recognition is the shame associated with the failure to meet performance targets. Ideally, financial incentives should consist of a mixture of rewards and penalties.

  Prizes are a form of positive reinforcement where providers know that there will be a reward for excellent performance, but they do not necessarily know in advance what standard will attract the reward, or what the value of the prize will be. Where commissioners are obliged to allocate a specific sum of money for a programme, the total value of the prize may be stated in advance, but how it will be shared among providers may only be known at the end of the assessment period, once relative performance has been ascertained.

  Prizes are widely used in the academic and scientific communities, but they were also used in Employment Zones, one of the earliest stages in the development of the UK government’s contracts for welfare to work. Since neither the performance target nor the amount of the bonus was disclosed by the commissioner, the prize acted as a ‘last mile incentive’, encouraging providers to find employment for as many jobseekers as possible.

- **Reputation:** Commissioners get a great deal of additional effort for free when providers believe that good performance will be recognised when contracts come up for renegotiation. While not originally designed with this in mind, prison inspections in the UK provide an outstanding example of reputational incentives
at work and there is strong anecdotal evidence that prison management firms worry about the tone as well as the content of these reports.

Reputational incentives involve the intermediation of a public official who is capable of exercising professional judgement in the interpretation of results. As a result, they are less vulnerable to gaming or distortion through extraneous influences than exclusive reliance on a set of mathematical formulae. Reputational incentives involve weighing and not just counting.

- **Regulation and certification**: In a modern society, there is a great deal of generic regulation, covering such issues as employee relations, health and safety, professional conduct and environmental protection, that contributes to the mix of incentives surrounding public service providers. The medical profession and the pharmaceutical industry, for example, are highly regulated, so that any use of payment-by-outcome in long-term condition management must have cognisance of this wider framework of constraints and incentives.

  This also applies in the field of criminal justice, where human rights legislation serves to limit the range of interventions that might be used. In the UK, a group of experts, organised as the Correctional Services Accreditation Panel, also assists the Ministry of Justice in the certification of new programmes for offenders.

  In a deep market where there are opportunities for repeat business, and where providers believe that reputation matters, companies will invest in securing third-party certification as a way of signalling to potential customers that they are different from their competitors. It is for this reason that public service providers pursue RoSPA (Royal Society for the Prevention of Accidents), BiTC (Business in the Community) or Public Servant of the Year awards and other ‘kitemarks’ of corporate social responsibility.

- **Norms and Culture**: Professional norms will also reduce the amount of gaming on the part of front-line service providers such as doctors, teachers and social workers. North American research found that public servants responsible for screening potential beneficiaries of job training schemes did not engage in creaming where that would have improved the financial rewards to their employer. Professional culture was an important factor in this case, although there was also a prohibition under the scheme on financial incentives being cascaded down to front-line staff. ⁴⁰
However, professional norms may constrain gaming even where individual providers do have the potential to capture some of the financial gains. It is generally considered that cultural incentives constrained self-seeking behaviour under GP fundholding in the UK during the early 1990s where, on the face of it, general practitioners might have been expected to restrict access to secondary care to generate savings.

- **Governance**: In some complex public-private partnerships, control is exercised not only through contractual instruments, but also through governance arrangements that are similar to the controls employed within traditional hierarchical organisations. Partnership boards (in the case of joint ventures), contract boards, and memoranda of understanding supplement the financial and other hard incentives built into the contract.

Thus, financial incentives are only one of the tools available to commissioners, and this report concludes that policymakers should look to the much wider range of instruments in the policymaker’s toolchest. This conclusion is similar to that arrived at by James C. Robinson in a 2001 article on the different incentive regimes in American medicine:

There are many mechanisms for paying physicians; some are good and some are bad. The three worst are fee-for-service, capitation, and salary. Fee-for-service rewards the provision of inappropriate services, the fraudulent upcoding of visits and procedures, and the churning of ‘ping-pong’ referrals among specialists. Capitation rewards the denial of appropriate services, the dumping of the chronically ill, and a narrow scope of practice that refers out every time-consuming patient. Salary undermines productivity, condones on-the-job leisure, and fosters a bureaucratic mentality in which every procedure is someone else’s problem. But American medicine exhibits numerous interesting compensation schemes that blend elements of retrospective and prospective payment, of fee-for-service, salary, and capitation. These innovations seek a middle ground between high- and low-intensity incentives, between piece rates and straight salary. Payment mechanisms also are embedded in and supported by nonprice mechanisms – i.e., by methods of monitoring and motivating appropriate behaviour that may have financial consequences but rely more directly on screening, socialisation, profiling, promotion, and practice ownership.
Managing Gaming

Gaming has already been mentioned a number of times this report, partly because it is such a pervasive problem in payment-by-results and partly because many of the tools discussed in this report are deployed to reduce the incidence or impact of harmful gaming. A fuller discussion is justified because, while the academic literature has recognised that some forms of gaming are beneficial, the policy debate has not yet caught up.

Adapting Carolyn Heinrich’s definition, gaming can be said to occur when providers increase their measured performance in ways that do not increase their performance in relation to the primary outcome sought. It can be difficult to avoid creating the perverse incentives that lead to gaming because primary outcomes can rarely be incentivised directly. Rewarding the achievement of a range of intermediate outcomes or high-level outputs, even where commissioners put an enormous amount of effort into selecting the right ones, still has the risk that providers can perform well on these while failing to deliver the primary outcome.

However, the risk of ‘gaming’ increases when commissioners have failed to identify all of the contextual outcomes of a programme. In so doing, they therefore fail to design a system that incentivises all of what they want the programme to achieve, and are then disappointed when providers fail to deliver results that are not specified and incentivised. While commissioners may perceive this as gaming, the term is inappropriate when applied to the actions of providers, since they could not have been expected to deliver what was not communicated to them in their contracts or the design of the system overall. This kind of misunderstanding can only be avoided if commissioners take seriously the initial task of selecting outcomes and agreeing detailed programme objectives.

Baron and Smith have argued that different kinds of innovation are likely to result from the introduction of performance incentives. Depending on the goals, some types of innovation that are sometimes labelled as ‘gaming’ could be beneficial to the programme as a whole.

• Change in the (technical) efficiency of service provision: This refers to improvements in the quantity and quality of the effort put forward by providers and is rarely, if ever, unwelcome. In fact, one of the reasons commissioners may choose to implement payment-by-outcome is that it encourages providers to make managerial changes that improve both efficiency and effectiveness.
• **Change in the persons served:** This kind of innovation is often labelled as gaming and attempts are made to regulate provision or incentivise providers so that they work with the same range of clients that were served prior to the introduction of payment-by-outcome. However, providers’ search for those individuals they judge to be most likely to respond to intervention may increase efficiency. Indeed, commissioners should exploit the information about the kinds of individuals providers are creaming and parking to segment the population differently, adjust incentive payments, introduce new measures and standards, and make other changes so that the system produces outcomes that are closer to those originally intended.

For example, creaming (where providers focus on those individuals closest to the labour market) and parking (where the most difficult-to-help clients are provided only a minimal service) are relatively common in welfare to work programmes. They occur because commissioners have not sufficiently incentivised providers to work with the most disadvantaged, for example by paying them more when they achieve outcomes for this group. Clearly, this may be a problem for reasons of equity, but creaming and parking also provide commissioners with valuable information about the actual cost of helping those jobseekers furthest from the labour market, empowering them to make the changes necessary to ensure this group receives the services they require.

• **Change in the mix of services provided to clients:** Again, changing the range of services clients are able to access could be considered gaming since commissioners may feel citizens are receiving a poorer quality service as a result. However, if payment-by-outcome has the effect of stimulating greater efficiency in service delivery, then we should expect to see a change in the mix of services.

These changes are only problematic insofar as commissioners have not adequately specified what they would like providers to achieve. For example, with the introduction of payment-by-outcome in welfare to work, many providers stopped sending jobseekers on lengthy skills courses and became much more work-focused. While less investment in human capital development might be considered undesirable, an approach that moves people into work more quickly and less expensively is more efficient, certainly in the short term. There is some evidence to suggest that a ‘work first’ approach such as this may also be more efficient over the long term.
• **Outright gaming:** Sometimes providers exploit design weaknesses, undertaking activities which allow them to perform better on the performance measures in ways that have little if any bearing on the primary outcome. This is outright gaming and commissioners should seek to reduce or eliminate the design features that allow it since it means providers are able to earn payments for non-welfare-improving activities. Some of the most outrageous examples of this come from the health sector, where studies have found that in order to perform well on waiting time targets, staff in some hospitals removed wheels from stretchers so they could be designated as beds and the patients redefined as having been admitted. In extreme cases, gaming may amount to fraud, and thus can be dealt with through the criminal law.

For the aforementioned reasons, it is important for commissioners to assess individual cases of gaming to determine whether or not they are truly undesirable. In general, gaming will be considered most harmful where it leaves certain groups of individuals in a worse position than prior to the introduction of payment-by-outcome. Where services under a payment-by-outcome scheme are additional, there can be fewer objections on these grounds. For this reason, it may be easier to pilot payment-by-outcome in additional services to allow learning to occur, before applying the model to core services.

Where gaming is undesirable, commissioners will need tools that help reduce or eliminate the perverse incentives that cause it. Adjusting the mix of performance measures so they accurately reflect the primary outcome sought; using measures appropriately, assessing impact and interposing discretion where necessary; changing the intensity and diversity of incentives; and segmenting the population differently (discussed in chapter 6) are among the tools commissioners can use to reduce harmful gaming.
In designing performance incentives in the workplace, firms do not rely solely on employee remuneration. To the contrary, a substantial proportion of workplace incentives are embedded in the design of different jobs. Much the same applies to the performance regimes constructed for motivating and controlling independent providers.

Apart from the design of the performance measures and the mix of financial and non-financial incentives, commissioners have several other instruments through which they can exercise control. These include the boundaries of the population whose outcomes providers are expected to improve, and the duration of the contract, elements that are similar in effect to job design within the firm.

**Population Segmentation**

It is inherent in the process of policy implementation, and particularly in programmes relying on payment-by-outcome, that commissioners will focus on a specific population whose welfare the policy is intended to improve. Thus, a job placement programme might focus on finding sustainable work for the *long-term unemployed*, or an offender management policy on accelerating desistance among *prolific offenders*. However, population segmentation can also be used as a tool for improving the efficiency and effectiveness of payment-by-outcome regimes, and it is with these applications that this chapter is concerned. Policymakers appear to segment their populations for several different reasons.

- *Different levels of need*: While all members of a particular population (say, the long-term unemployed) may have the same ultimate need (in that case, sustainable employment), the reasons why they are not able to realise this
ambition may fundamentally differ, and some groups of individuals may be much further from achieving it than others.

Different kinds of jobseeker will have different needs, and it is possible that for some individuals, the best policy intervention may not be as suited to resolution through payment-by-outcome. For jobseekers close to the labour market, performance measures based on entry into the workforce and holding down a sustainable job may be entirely appropriate; however, for more disadvantaged jobseekers with complex needs, solutions based on building human capital over time may need to be designed in a different way.

In the case of offender management, policymakers might focus a particular programme on sex offenders, or tailor services to prisoners with mental health problems. While payment-by-outcome may still be appropriate in these cases, commissioners may decide that the needs of different groups are better addressed through separate schemes with different architectures.

- **Costs and benefits:** In some cases, policymakers may choose to focus a policy intervention on a particular sub-group for the explicit reason that the costs of delivering the outcome to these clients are expected to be lower, deliberately targeting the low-hanging fruit. For example, commissioners may concentrate on older male prisoners who are more inclined to desist from criminality for reasons of maturation and rising opportunity costs. Alternatively, they might focus on prolific offenders because of the greater magnitude of the economic gains expected from increased desistance among this cohort.

- **Reduced gaming:** One of the ways that commissioners can reduce the scope for creaming is to segment the population into several different classes of beneficiary that are internally homogeneous, and to pay providers different fees based on the difficulty of delivering outcomes to these groups. Job Services Australia has used the Job Seeker Classification Instrument to separate clients into three (sometimes two) different groups based on disadvantage, each attracting a different level of payment.

In the UK, the population of jobseekers used to be segmented according to benefit type, presumably on the assumption that this might indicate fundamentally different classes of need. However, benefit types were not sufficiently useful in categorising clients for the purposes of job placement, and the Department for Work and Pensions is moving to a classification tool similar to that used in Australia.
• **Measuring performance:** Where they expect to rely on statistical comparisons or yardstick competition to assess performance, commissioners will need to consider the homogeneity of populations to ensure that comparisons are meaningful.

• **Increasing price competition:** Where contracts are being awarded through competitive tender, homogeneity also makes it easier for commissioners to intensify price competition. Much the same occurs at auctions where auctioneers bring together lots of broadly similar items – competition tends to be more vigorous than when a miscellaneous assortment of items is put to auction as a job lot.

• **Extraneous variables:** The recently-published Criminal Justice Green Paper recognises that population design is also influenced in part by the provider’s control of extraneous variables. In such cases, population selection will have an impact on the kind of provider that is invited to participate in the scheme.

  For example, in addressing offenders who are sentenced to more than 12 months in prison, the report notes that they will typically spend time in a number of different prisons, receiving rehabilitative interventions from a variety of different providers. As a result, ‘It can be difficult to identify which organisation deserves the most credit for a reduction in re-offending.’ The recommended solution is to target the proposed payment-by-results system at the providers of probation services, since they ‘have the most continuous contact with the offender, from the sentence to completion’. By contrast, more than half of prisoners serving sentences of less than 12 months spend their time in a single prison, and the Green Paper argues for incentives aimed at reducing re-offending to be directed to prison providers.\(^{45}\)

**Dynamic Design**

It is sometimes assumed that population design must take place in advance of procurement, with boundaries remaining relatively static throughout the life of the contract. However, with a learning system, it is possible for segmentation to be dynamic, with adjustment and adaptation over time.

On this view, the ideal segmentation of the population cannot be known in advance, but will be discovered through exploration and systemic learning. Thus cream-skimming in performance contracting is seen not only as inevitable, but – assuming systems are designed well and there is scope for adaptation on the part of commissioners – as a necessary part of the learning process.
This is often difficult for policymakers to accept, since cream-skimming has traditionally been viewed as opportunistic behaviour to avoid delivering services to the population for which the contracts were let. It is politically embarrassing, since providers profit from delivering something less than what was contractually committed.

However, if commissioners do not fully understand the boundaries of different population segments, and if there is previously unidentified ‘low-hanging fruit’, for whom outcomes can be delivered at significantly lower cost, then providers make an important contribution to the efficiency of the programme overall by identifying these individuals or classes of individual. A political problem arises when commissioners are locked into long-term contracts so that they are paying more than necessary for an extended period of time. On the other hand, where commissioners create an adaptive system, so that exploration is encouraged and lessons are learned, the political risks can be significantly reduced.

An obsession with not paying for ‘deadweight costs’ – the costs associated with assisting people who would have found a job or given up re-offending on their own – reflects a static view of population design, since it assumes that this group can be known in advance.

The prospectus for the UK government’s proposed new ‘Work Programme’, released in November 2010, deals with the question of differential pricing skewed in favour of more difficult customers. The prospectus was not supportive of the so-called accelerator funding model which would allocate funding at an increasing rate as more clients were placed in employment. This model was rejected on the basis that ‘it encourages providers to support the easiest to help into work first’ and the paper opted for a system of classifying job seekers up front, which it assumed was more equitable.46

Whilst not endorsing the accelerator funding model, this report argues that there are strong public benefits in identifying the low-hanging fruit, and assisting this segment of the job-seeking population as soon as possible. A problem only arises if commissioners are obliged to pay unacceptably high prices for identifying this group.

Static classification models may not be sufficiently responsive to alternative (and possibly more effective) ways of classifying beneficiaries. Providers may well have access to local knowledge that will enable them to target resources much more effectively on the relevant segment of the population. There is some evidence
that alternative classification systems have emerged under Flexible New Deal, as providers have organised themselves and their supply chains in fundamentally different ways.47

Dynamic design is evident in some of the risk-sharing schemes being employed in the roll-out of innovative new pharmaceuticals. For example, in 1999, the North Staffordshire Health Authority, in collaboration with Pfizer, introduced a pilot scheme based on an outcome guarantee for a new statin (a drug used in lowering cholesterol). Pfizer agreed to refund all costs of the drug if the expected reduction in cholesterol levels was not delivered. This gave Pfizer an incentive to identify as early as possible, those patients for whom the drug was unlikely to make a difference, so as to exclude them from the population being served and thereby reduce its financial exposure, and the company cooperated with medical practitioners to identify the target group.

Other outcome-based agreements in the pharmaceutical sector monitor results on an ongoing basis and rely on ‘continuation criteria’ to ascertain whether treatment should be extended. In Australia, a registry was maintained for the administration of bosentan to patients enrolled for the treatment of pulmonary arterial hypertension, with strict criteria for excluding those who were assessed to be not responding.

In both these cases, the system was designed from the outset to identify as soon as possible those patients for whom the interventions were not appropriate. There was no suggestion that the drugs should be withheld until a solution had been found that worked for all sufferers of the diseases in question. Payment-by-outcome was deliberately employed to segment the population, and delivered targeted interventions to those for whom they would work best.

Much the same approach should be adopted in payment-by-outcome for public services such as offender management and welfare to work. Commissioners should not expect that in the beginning, they will understand all the salient characteristics of the population of potential beneficiaries. Because of their detailed knowledge of local conditions, front-line public servants always play a significant role in the making of policy, and the highly devolved and adaptive nature of decision-making under payment-by-outcome means that the influence of providers will be even greater.48 Rather than trying to ignore these qualities of payment-by-outcome, commissioners should make them work to their advantage.
Contract Duration

Commissioners’ ability to adjust the terms of the performance regime as they learn will depend in part on the length of the contracts. Shorter contracts will enable commissioners to review the terms frequently, reducing prices to capture efficiency savings, segmenting the population differently, and generally adapting the contract to suit changing circumstances. Longer contracts make adjustment of this kind more difficult, but they have other benefits that may outweigh the loss of flexibility.

Thus the length of contract ultimately selected will depend on several variables. One is the length of time it takes to deliver the outcome or the evidence of major milestones and this will differ from case to case depending on the nature of the service in question. Of course, contract length will also be determined by the duration of the performance period. For example, it takes a certain amount of time to prepare jobseekers to return to work, and commissioners must take this into account in the design of contracts. In addition the Department for Work and Pensions imposes a 13- or 26-week measurement period on providers once jobseekers are in work, and under the Work Programme, they are proposing measurement periods of up to two and a half years. Contract duration must be at least as long as it takes to produce and measure one outcome or proxy output, although commissioners may wish to allow several cycles of outcomes to be delivered before re-competing the service.

Second, some services may require greater initial investment than others and therefore the supply side will demand longer contracts to ensure they reap the returns. Thus, we should expect contract duration to be longer under invest-to-save schemes and wherever greater performance risk is shifted to providers.

Commissioners may also wish to let longer contracts in order to stimulate innovation in service delivery. In public service markets, where competitive tendering is highly transparent and where commissioners are reluctant to recognise intellectual property rights in major innovations, the duration of the contract may be one of the few protections that providers have to secure a return on transformational initiatives that require some investment in research and development.

On the other hand, very long contracts may weaken the competitive pressure on providers to deliver better outcomes, thereby reducing the level of innovation and service quality. Longer contracts have the tendency to thin the market, since fewer tenders mean less frequent competition and fewer opportunities for new providers to enter the market. Commissioners need to consider how to keep the threat of contestability real in order to maintain a healthy market.
Shorter-term contracts and more frequent competitions are one way of doing this, although as noted above, they bring with them their own disadvantages. Contract scale – the size of the population being served – is another variable that commissioners can adjust to increase market depth and encourage new entrants, as well as staggering the dates when contracts are thrown open to tender, so that providers face an ongoing series of opportunities to tender their services and demonstrate the value of their reputation.

Where longer contracts are preferable, commissioners can impose minimum performance standards which allow them to terminate contracts in the case of unsatisfactory performance. The Department for Work and Pensions has made this a feature of the Work Programme: providers will be given five-year contracts but must deliver a minimum level of employment outcomes for certain categories of jobseeker; failing to do so will result in contractual action, including contract termination if performance does not improve.49

Another interesting innovation from Australia is the rolling local area tender. At the end of the third Job Network contract in 2006, all contracts of providers performing to a satisfactory standard were renewed for another three years. However, six-monthly reviews allowed commissioners to replace poor performers throughout the period of the extension, thereby maintaining pressure on providers to continue to deliver high quality services.50

Minimum standards, automatic contract renewal and rolling tenders allow commissioners to avoid re-competing services where performance is good and to replace poor performers where necessary. The administered contract may be another option. In this case, an independent arbiter or regulator is appointed whose role is to re-set prices at pre-determined intervals throughout the contract term. This arrangement was used for the London Underground PPPs, where commissioner and contractors were obliged to commit themselves to long-term arrangements with only limited understanding of the challenges the contractors would be required to address. The Department for Work and Pensions has contemplated although not yet implemented the use of such a model in the welfare to work market in the UK. Of course, there are limitations with such arrangements, but they have proved relatively robust in the utility sector for more than a century in North America and the UK.51
System Design

If the term ‘commissioning’ means something more than just ‘procurement’, then it must include a contribution to the selection and design of the overall systems within which implementation takes place, as well as the design of the policy framework within which procurement occurs.

Ownership of the Residual

After the decision is taken to commission services from independent or semi-independent providers, policymakers must still decide whether they will be procured through choice-based markets, where individuals or their agents negotiate directly with providers, or competitively-tendered markets where government agencies procure services on behalf of their clients. This is not just a matter of ideological preference, but will be heavily influenced by the nature of the service in question and the challenges associated with different design characteristics.

This report is concerned with public services that are competitively-tendered, rather than choice-based systems. However, as noted in the previous chapter, one of the challenges associated with commissioned services are the perverse incentives that arise at the end of the contract period. In the case of long-term condition management in the UK, the health of the population at the end of the contract period is a valuable asset, but it is owned by the commissioner and not the provider, and unless this asset can be valued and the provider compensated, there will be an incentive to under-invest in the long-term health of service beneficiaries.

This perverse incentive can be ameliorated to some extent by extending the contract period, and/or by confining the population to patients with particularly severe conditions, but as long as there is a commissioner-provider split, this problem will remain.
Choice-based selection is capable of overcoming this problem, particularly where patients (or agents on their behalf) exercise that choice in selecting their health insurer, rather than just selecting their health provider. In this case, the insurer owns the residual asset, and thus it is in their interest to invest in the management of long-term conditions. That it does operate as a powerful incentive for health insurance firms is evident from the fact that the great pioneers in long-term condition management have been the managed care organisations of North America.

The other benefit of a choice-based system is that service contracts are renegotiated periodically, thereby avoiding the threshold effect that accompanies the termination and re-tendering of a single contract for a large population of beneficiaries. This distortion might be ameliorated to some extent if there was the prospect of the contract being rolled over in the case of good performance, but the competition norm in public contracting is so strong that governments are generally unwilling to extend contracts in this way.

It is possible that choice-based markets may also stimulate a greater investment in research and development. One of the disadvantages of commissioning public services through competitive tendering is that intellectual capital is undervalued. The public nature of the tendering process means that new ideas are quickly disseminated across the system as a whole. Of course, this also has significant benefits, but it does have the effect of discouraging significant investment in long-term research and development. Choice-based systems are not as transparent as those based on competitive tendering, and this might also help to explain the significant advancements made in recent decades in the managed care organisations of North America.

Of course, individual choice is not possible in certain public services. Offender management is the obvious example – it is difficult to imagine how one might construct a corrections system based on prison vouchers. And there are many other challenges associated with relying on individual choice in public services that may tip the balance in favour of commissioning services through a public agency. It is beyond the scope of this report to resolve this particular issue; however, it is important to recognise that alternative market models may overcome some of the difficulties associated with contracting for outcomes.

Building a Learning System
The most complex payment-by-outcome systems currently in operation – those employed in assisting the long-term unemployed into jobs – have been under development for more than 60 years. The earliest example of performance incentives
being used for staff involved in job placement dates to 1948: this was a system for the management of full-time employees, and the incentives were reputational rather than financial, but it bore many of the design elements of a modern payment-by-results contract.\textsuperscript{52}

Financial incentives and performance contracts were first introduced in some of the US states in the 1970s, and the federal government passed the Job Training Partnership Act in 1982, a system that has been amended and improved in the decades since. In Australia and in the United Kingdom, jurisdictions that have explored payment-by-outcome models since the mid-1990s, there has been a succession of schemes as lessons have been learned and confidence has developed in moving to higher-level objectives and transferring more risk.

Policymakers in other service areas are now studying these welfare to work models in an attempt to shorten the development process; however, there are significant differences between public services, and only rarely can designs be borrowed directly.

One of the major conclusions to be drawn from this project is the desirability of payment-by-outcome systems being developed over time. Any move to strong performance incentives and outcome specification will inevitably be a process of discovery, with initial mistakes and misunderstandings. Academic economists have argued the advantages of ‘adaptive contracting’, where commissioners deliberately leave contracts incomplete, using contract renegotiation to learn from the opportunistic behaviour of providers.\textsuperscript{53}

By adopting this approach in payment-by-outcome contracting, commissioners should be able to exploit gaming behaviour such as parking and creaming, to identify previously unidentified segments of the beneficiary population and redesign the contracts so that they better serve the public interest.

\section*{Designing Procurements}

Since this report is based on the assumption that payment-by-outcome is implemented through a system of performance contracting, it follows that some consideration must be given to the design of the procurement regime through which providers are selected.

Competitive tenders are intended to reveal the providers who can deliver the best value for money – the greatest improvement in service quality for a fixed unit of cost; the greatest reduction in spending for a fixed unit of service; and/or the greatest capacity for managing performance risk in the outcome in question.
This is challenging enough where commissioners and providers understand the nature of the service and how it should be priced, but there is a grave risk of generating a ‘winner’s curse’ in a price-based competition where commissioners and providers do not agree upon the fundamental characteristics of the service being put to tender. Economists have defined a ‘winner’s curse’ as a procurement where the winner inevitably bids a price/service package that it is not able to deliver or assumes risk that it is not capable of managing.

Particularly in the early stages, payment-by-outcome procurements seem to bear many of the characteristics of the winner’s curse, and as discussed in the following chapter, this is what appears to have happened in the UK in the competition for Pathways to Work.

One way of overcoming this difficulty is to organise the competition so that it is focused on quality rather than price, although the risk in this case is that commissioners may pay too much for services, resulting in large profits and poor value-for-money for taxpayers. This appears to be what happened when the Australian government procured its first outcome-based contract for the Working Nation programme in 1994.

In an attempt to resolve the problem without creating a winner’s curse, Australia set a minimum price which potential providers could bid above but not below. This did not appear to work, however, since providers’ bids did not vary much from the floor price. One proposal that may enable commissioners to be confident of both service quality and value-for-money is the imposition of contracts with periodic price re-sets that maintain pressure on providers to deliver efficiency savings and allow these to be captured by commissioners.

Ultimately, decisions about how to procure services must take into account a large number of factors, including the level of innovation commissioners expect to see at the bidding stage and thus how easy it will be to compare the quality and cost of bids; the relative importance of service quality versus efficiency savings; the level of experience of providers of delivering services; and assessments of providers’ ability to cost services accurately.

It is possible that where payment-by-outcome is implemented for the first time, fixed-price competition based entirely on quality may be more appropriate, whereas in mature markets commissioners can be more confident that price-based competition will not lead to a winner’s curse.
Case Study 1: Welfare to Work

Performance management systems have been used by US governments in their welfare to work programmes since the late 1940s, although the first large-scale, federally-funded programme that used performance standards, the Job Training Partnership Act (JTPA), was rolled out in 1982. Since then, two other federally-funded initiatives and a variety of state programmes have been implemented. Australia began to experiment with payment for outcomes in its Working Nation programme in 1994. Working Nation was replaced by the Job Network in 1996, and in 2009 the Job Network evolved into Job Services Australia. The UK’s first payment-by-outcome programme, Employment Zones, was rolled out in 2000. Subsequent programmes have built on the New Deal, another welfare to work programme, rather than the Employment Zones model, as government has sought to increase the levels of risk transferred to providers.

Although these programmes all paid for outcomes, the differences among them are striking. In the American JTPA programme, the primary outcome specified in the legislation was to maximise the ‘return on investment in human capital’, where human capital was defined as a ‘set of skills and knowledge’ acquired through education and training. The population targeted for the intervention suggests poverty reduction was another implicit objective. In Australia and the UK, the focus has been more explicitly on ensuring benefits claimants find employment, promoting social inclusion and reducing the cost of unemployment to government.

The measures selected to determine whether these objectives were achieved reflected the variation in primary outcomes. Under JTPA, clients’ average wage on entering employment was one of four measures used to determine an agency’s eligibility for an outcome award. The US is the only one of the countries studied here that placed emphasis on increasing jobseekers’ earnings, used as a proxy
for increased human capital. Australian and UK programmes have focused on measures of sustainable employment, namely the number of jobseekers achieving 13 and 26 weeks continuously in work.

Incentive systems also differ. In the American JTPA, states had discretion over the ways in which eligible agencies were rewarded. In most states, agencies simply had to meet a certain standard of outcome achievement to receive an award. However, some states gave the entire bonus to the best-performing agency while some divided it among all agencies performing above a certain threshold. Other states varied the award according to performance relative to the standards, so agencies which far surpassed their targets received more than those which simply complied.57

The Employment Zones programme in the UK was quite different. Providers received some funding upfront, a proportion of which they were entitled to retain if jobseekers found work within a certain period of time. Providers also received standard payments for jobseekers placed in employment and those who retained work for 13 weeks. Subsequent UK programmes such as Pathways to Work and the Flexible New Deal have similar incentive systems to Australia’s Job Networks, with providers receiving service fees of varying levels for accepting jobseekers onto their programmes and then earning outcome payments when jobseekers retain jobs for 13 and 26 weeks. A key difference between these two countries is that Australian programmes have also tended to measure and reward education and training outcomes, while in the UK the incentives have steered providers to focus much more heavily on employment outcomes.58

Finally, the client groups targeted and the ways in which outsourcing has been used to serve different groups of jobseekers varied quite dramatically. Certain American programmes have been aimed at individuals living below the poverty line who do not have to be unemployed to qualify. In Australia, jobseekers are classified according to their level of disadvantage, and this determines their eligibility for services. Those categorised as closest to the labour market are eligible only for job matching services, while the most disadvantaged qualify for intensive assistance.59

The private and voluntary sectors play a role in providing services for all clients. In the UK, eligibility for programmes has historically been determined by length of unemployment and the type of benefit claimed. In general, those unemployed for less than twelve months qualify for services from Jobcentre Plus, the public employment service, while the long-term unemployed are served by private and
voluntary sector providers specialising in personalised case management, with a separate programme for those with disabilities. The Coalition Government has plans to establish a Work Programme for most jobseekers regardless of benefit type, and certain jobseekers (for example those classified as ‘facing significant disadvantage’) will have access to this programme after only three months of unemployment.60

This brief chronology demonstrates that payment-by-outcome can be applied to welfare to work programmes in different ways. Each programme has a specific context which affects its design and level of success. The following sections analyse those tools that are most critical to the successful use of payment-by-outcome in welfare to work schemes.

Population Segmentation
The population of jobseekers in any country is diverse. They face different types of barriers to work and some are further from the labour market than others. Working with a diverse population has implications for the design of a payment-by-outcome system. First, although the primary outcome for all jobseekers may be finding sustainable employment, the measures used to assess progress may depend on how far a group is from the labour market. Second, providers may need to be paid more to help highly disadvantaged jobseekers who will cost more to get into work. Finally, the different kinds of jobseekers with which providers work may make it difficult to compare their performance, so commissioners may find it necessary to segment the population into groups to make measuring providers’ impacts feasible.

Selecting Measures
The UK, the US and Australia all desire one or two primary outcomes for all clients. However, applying the same measures of progress to all jobseekers could encourage gaming and reduce providers’ flexibility in delivering services tailored to their clients’ needs. Therefore commissioners may need to create more than one welfare to work programme, so that providers of different programmes are incentivised to achieve different measures, reflecting the group of jobseekers with which they work.

If all jobseekers are served by the same programme, and that programme measures only employment outcomes, there is a risk that providers will park disadvantaged jobseekers who have complex, non-vocational barriers to work. For jobseekers closest to the labour market, measures of entry into the workforce and retention of employment are appropriate. However, for more disadvantaged
jobseekers, for whom employment outcomes are unlikely in the short to medium term, a separate programme that incentivises providers to improve education and skills outcomes may well be necessary.

Measuring employment outcomes also narrows the scope of interventions providers are likely to offer. Providers may feel obliged to adopt a ‘work first’ approach which may not successfully address the needs of very disadvantaged clients. Where education and skills outcomes are rewarded, providers are much more likely to invest in developing the human capital of their clients. Commissioners can segment the population based on jobseekers’ distance from the labour market, referring groups to different programmes that define and measure outcomes in different ways.

Australia has adopted this approach. Thus, providers under the Personal Support Programme operated separately from the Job Network and under a different set of incentives which, while still rewarding employment outcomes, focused providers much more on developing human capital through education and skills courses.61

There is some evidence that highly disadvantaged jobseekers benefit more from employment-focused than human capital-focused programmes. However, there is also some evidence that the most successful approach is one in which providers personalise clients’ back-to-work plans so that some are immediately work-focused while others begin with basic skills training or other activities. Programmes that exclusively measure and reward employment outcomes may prevent providers from being flexible about which activities clients undertake first. Therefore it may be appropriate to refer to other programmes those jobseekers who are likely to benefit from non-vocational support first.62

Setting Prices

To reduce incentives to cream and park, commissioners may pay providers more for securing outcomes for highly disadvantaged jobseekers; however this demands a reliable and trusted means of classification. In Australia, the Job Seeker Classification Instrument is used to categorise clients into three different groups, each attracting a different level of payment. Providers have the option of petitioning to reclassify jobseekers they feel have been miscategorised.63 The instrument is only accurate enough to classify jobseekers into broad categories, so that the groups are still quite heterogeneous. While imperfect, it is likely this does reduce the incidence of parking.
Commissioners may also choose to establish the proportion of jobseekers that fall into each price category and increase the level of payment to providers as more clients are served. For example, commissioners might not pay anything for the first 10% of employment outcomes, on the assumption that these jobseekers would have found work without intervention, £600 per outcome for the next 15% of jobseekers, £1,000 for the following 15% and so on. A target accelerator model such as this would eliminate the risk of misclassifying individual jobseekers, but it would still rely on commissioners correctly identifying the proportion of jobseekers in each category and the approximate cost of getting them into work.

**Measuring Impact**

Because the population of jobseekers is diverse, it is likely that providers will serve different proportions of easy and difficult clients. This makes comparing the performance of different providers, a common way of distinguishing the impact of the service from that of external variables, very difficult.

Measuring raw outputs – the numbers of jobseekers retaining employment for 13 and 26 weeks – is relatively simple. However, in order to measure the impact of providers on those outcomes, it is necessary to control for external factors that may have impacted on outcomes. This can be achieved in various ways. In general, commissioners of welfare to work programmes have not been able to compare programme results with those of a control group. An alternative way of attributing impact lies in using a statistically derived comparator, based on the results that commissioners would expect providers to achieve, taking into account jobseekers’ characteristics and labour market conditions. However, this relies on the statistical model being robust.

One common way of attributing impact is therefore to compare the results of two providers operating in the same conditions, referred to as yardstick competition. Providers operate in the same labour market and serve similar populations of jobseekers, achieved through referring similar proportions of easy and difficult to help jobseekers. Yardstick competition reveals information about the impact of external variables since if two or more providers are underperforming in relation to expectations, there is a prima facie case that difficult economic conditions have impacted on outcome achievement. Alternatively a large discrepancy between the results of providers is an indication of underperformance by one.
Collusion, often raised as a problem with this model, does not seem to have been an issue in the UK, where several Employment Zones and Flexible New Deal districts operated with yardstick competition. This is possibly due to the shadow of the future, and the importance providers place on protecting their brands.

Another challenge lies in classifying jobseekers accurately so that providers serve similar client mixes. The tool used to classify jobseekers must be robust, taking into account the key variables that may impact on an individual’s ability to obtain and retain work. Alternatively, providers can serve different types of jobseekers and commissioners can adjust for this statistically before making comparisons. However, statistically controlling for client mix still requires agreement on the characteristics that make jobseekers more difficult to help. Australia makes no attempt to ensure providers serve similar mixes of clients, but rather uses information about the jobseekers served over a certain period to make statistical adjustments to the number of outcomes achieved. This forms the basis for ‘star ratings’, the principal method of comparing providers.66

Designing Incentives
The design of incentives is critical in the creation of a payment-by-outcome system since they shape providers’ behaviour, sometimes in unanticipated and unacceptable ways. The principal rationale for paying for outcomes is that it aligns providers’ interests with those of the commissioner. However, since many primary outcomes are not directly measurable, commissioners must incentivise the achievement of intermediate outcomes or outputs instead. Problematically, these are only able to capture some of the aspects of the primary outcome, which means that providers may have a perverse incentive to increase their measured performance rather than their actual performance in relation to the ultimate outcome. Commissioners have experimented with scaled payments, retrospective adjustments, and prizes, all of which are intended to limit the perverse incentives created by paying for outputs.

Paying for Outputs
Thirteen- and 26-week employment measures do not capture the totality of what commissioners want providers to achieve. These outputs do not, for example, specify which kinds of jobseekers commissioners want providers to return to work or give a detailed indication of the quality of employment providers must help jobseekers find.67 These omissions create opportunities for providers to game the system, for
example by focusing on jobseekers with few barriers to work (creaming), providing minimal services to very disadvantaged jobseekers (parking) or by encouraging jobseekers to accept jobs with little opportunity for advancement (an example of shading service quality). Commissioners can reduce the likelihood that providers will game by employing other incentives.

**Scaled Payments**
Increasing the amount of payment providers can earn for helping highly disadvantaged jobseekers into work should decrease the incidence of parking. Under Australia’s Job Network, jobseekers were categorised according to their distance from the labour market so that providers earned more for achieving outcomes for disadvantaged jobseekers. Since 2006, experts in the UK have been advocating a form of ‘target accelerator’ in which commissioners pay progressively more the more jobseekers providers return to work. But such a system has not yet been implemented and the Department of Work and Pensions has indicated that it is not supportive of this approach.

**Retrospective Adjustment**
Rather than estimating in advance the proportions of different types of jobseekers providers will serve and the likely labour market conditions in which they will operate, retrospective adjustment involves paying providers based on actual clients served and conditions experienced. While scaled payments reduce incentives to park, retrospective adjustment also takes into account the impact of labour market conditions on outcomes. As the adjustment must be made retrospectively, commissioners can either pay providers an expected rate and then request a refund or arrange additional payment to providers, or wait until the end of a measurement period to make payments. To the authors’ knowledge, no payment-by-outcome systems are currently using retrospective adjustment, although programmes such as the JTPA have adjusted based on predictions of expected performance.

**Prizes**
Prizes can act as ‘last mile incentives’ for providers to continue to achieve outcomes beyond the level where the cost of helping each jobseeker begins to outweigh the amount of the outcome payments. In Employment Zones, providers could earn a bonus for placing a certain proportion of their jobseekers in employment. The
amount of the bonus and the proportion of jobseekers providers were required to place in order to earn the prize were not disclosed. This ensured that providers had an incentive to find employment for as many jobseekers as possible, rather than simply finding work for those whom it was profitable to place based on the amount of the outcome payments.

The Procurement Process
The way in which services are procured can have a large and lasting impact on their quality and the outcomes they achieve. The long history of contracting welfare to work programmes enables analysis of the effects of the procurement process on payment-by-outcome systems.

An important lesson from the procurement of Provider-led Pathways to Work in the UK is that competition based heavily on price or risk transfer can lead to a winner’s curse where the winner inevitably bids an uneconomic solution. This occurs where price-based competition is fierce and there is uncertainty about the population to be served and the exact nature of the service to be delivered. Pathways bids were in large part assessed based on the numbers of outcomes providers pledged to achieve. Many potential providers based their estimates of the outcomes they could achieve on their experience delivering the voluntary New Deal for Disabled People programme. However, jobseekers on the mandatory Pathways programme faced complex barriers to work and were far less job-ready than providers had anticipated. Thus the cost to providers of helping these individuals was higher than expected.

This problem was compounded by the highly competitive bidding process. Providers were required to bid based on the number of jobseekers they would get into work, and this determined the unit price per outcome achieved. Partly because they had misjudged the population and partly because the process was so competitive, providers were overly optimistic about the numbers of jobseekers they would be able to get into work, which resulted in their unit prices being very low. Subsequently, providers were unable to supply the numbers of outcomes anticipated, and thus faced severe cash-flow problems. These two issues combined to help create an unviable supply market providing a ‘bare bones’ service to clients.

Australia provides alternative examples of how to procure services under payment-by-outcome. For the Working Nation programme (1994), the government opted to conduct a competition based on quality, since the market was new and
providers inexperienced. Running a competition based on quality is likely to encourage more innovation by providers who cannot make their proposals stand out based on a very low priced bid, but must ensure instead that their service delivery model is original. Moreover, by fixing the price, rather than allowing it to emerge through the procurement process, the commissioner assumed the risk of setting it too low so that service quality suffered or setting it too high, so that providers made large profits. This may have been a sensible risk, since at this early stage the Department of Employment and Workplace Relations, which was procuring the service, probably possessed more knowledge about costs than the inexperienced providers. When the next government scrapped the programme, it argued Working Nation had been expensive and ineffective at placing jobseekers in permanent employment. It is possible that the high programme cost was partly a result of the government setting the fixed price too high, and the generous profits made by some providers seem to confirm this.

The three Job Network contracts that followed were procured slightly differently. For the first contract, Job Network providers could bid to deliver up to three different services: job matching, job search training and intensive assistance. Providers could compete on price for job matching and job search training functions, which were fairly standard services. Bids were assessed for quality and then ranked by price. The Department was not required to accept the lowest bid for any service, but could instead trade off aspects of quality and price. However, prices for intensive assistance were fixed and providers competed solely on quality for this particular service. This was due to ‘concerns that the bidders would initially lack the expertise to cost the new service’.

For the second round of contracts, the Department decided to allow some price competition for the intensive assistance service. A 75% weighting was given to the quality of the services in the bid and a 25% weighting to price. The Department set a floor price which potential providers could bid above if they pledged to deliver ‘greater outcomes than the average’. This kind of competition was intended ‘to ease the transition to a fully competitive market for Intensive Assistance, and as a safeguard to protect service quality and reduce the risk of market failure’. In practice the floor set the price, as potential providers bid down to that level: ‘the difference between the average and minimum price for upfront payments for level B clients in [intensive assistance] was less than 7 per cent, compared with nearly 90 per cent for [job search training] (where no floor price was set)’.
The third round of procurement saw a return to fixed-price competition as the Department attempted to weaken the incentive to park by ensuring a sufficiently high level of compensation for helping the most disadvantaged.82

Thus commissioners can use the procurement process as a tool to promote innovation by selecting providers offering different models of service delivery, and as a means to secure efficiencies through price-based competition. However, where providers are inexperienced and there are uncertainties about the cost of delivering outcomes and the level of outcome achievement possible, using competition to lower the price could lead to a winner’s curse. In immature markets, fixed-price competition may be the most appropriate procurement method. To secure efficiencies with fixed-price competition, commissioners can implement periodic price re-sets to ensure they capture the efficiency savings made by providers.
Case Study 2: Offender Management

Despite recent decreases in re-offending rates, around half of adult prisoners in England and Wales are reconvicted within 12 months of release. The economic costs are considerable: in 2002 the Social Exclusion Unit estimated that imposing a custodial sentence at a crown court cost the criminal justice system £30,500 and incarcerating an offender cost an additional £37,500 per year. Reducing the rate of re-offending would therefore help the Ministry of Justice deliver the substantial cost savings required by its budget settlement in the 2010 Comprehensive Spending Review.

This case study considers the tools that commissioners of offender management services might use to implement a payment-by-outcome system successfully. In particular, it examines how performance can be measured and how using different measures may affect the scope for innovation.

Background

Following the Carter Review in 2003, the National Offender Management Service (NOMS) was established to bring together the headquarters of the prison and probation services. The intention was to give each offender a case manager who would follow the offender throughout imprisonment and probation to ensure continuity of relationships and service provision. NOMS, now an executive agency of the Ministry of Justice, is responsible for commissioning and delivering adult offender management services including the safe incarceration of offenders and rehabilitation services in custody and the community.

core focus was on the resettlement of ex-prisoners through the provision of key services, including accommodation; education, training and employment; mental and physical health; drugs and alcohol; finance, benefit and debt; children and families of offenders; and attitudes, thinking and behaviour. These are known as the seven ‘pathways’ to reducing re-offending. A variety of programmes designed to assist offenders to progress along these seven pathways are provided in prisons and in the community. The specific needs of each offender are assessed and offenders are referred to specialist services for support to meet their needs.

More recently, the government published its Green Paper, Breaking the Cycle, detailing further reforms to the criminal justice system, including plans for more effective rehabilitation and the use of payment-by-results to reduce re-offending.

Payment-by-Outcome Pilots
Since the Carter Review there has been growing interest in the use of contracting to drive a greater focus on rehabilitation and resettlement, with a variety of pilots directed to learning how payment-by-results might be implemented in this field. Some of the pilots focus on getting offenders into stable employment. Given the aim of reducing re-offending and the evidence that improving the employment prospects of offenders contributes to a reduction in re-offending, this is an output; however, the Department for Work and Pensions considers it an outcome.

While not yet based on payment-by-outcome, Path2Work was an early pilot directed to the improvement of employment outcomes that achieved comparatively good results. Managed by an alliance of private and voluntary sector providers, the scheme provided services to increase the employability of ex-offenders in the East of England from 2006 to 2009. It was voluntary and additional to other welfare to work services, including New Deal and Pathways to Work. Path2Work provided a variety of services including improving basic skills, job matching and advice on disclosing information about criminal convictions to employers. Offenders were also offered in-work support once they had found employment. Four- and 12-week employment outcomes of participants were monitored, and the results were promising. An evaluation carried out by Deloitte indicated that while Path2Work did not meet its targets, 30% of participants were placed in work, compared with 6% to 11% of offenders participating in similar programmes.

In September 2010, the Ministry of Justice announced a six-year pilot scheme catering for 3,000 male prisoners at Peterborough Prison serving sentences of less
than 12 months (a class of offender who would not normally receive post-release support). The contract with not-for-profit providers is based on payment-by-outcome, with the initial investment financed by a ‘social impact bond’ where third-party investors are compensated if social outcomes are achieved. Investors will be paid for reductions in conviction events compared with a control group.89 Payments start when the reconviction rate of the intervention group is 7.5% less than that of the control group, with increasing returns up to a maximum rate of 13%.90 The Peterborough pilot is the first in the world where private investors have assumed financial risk for reducing re-offending.91

*Job Deal* aims to help young people not in education, employment or training, prisoners with less than three years to serve, and offenders on community sentences into employment. It is part of a larger scheme funded by the European Social Fund and the Department for Work and Pensions, and is managed by NOMS. Phase One of the pilot began in 2010 and will run for two years.

The provider assigns each offender a case manager. Together they develop a tailored action plan and identify any specialist support the offender may need. The scheme is voluntary, although participants must commit to meeting with their designated case managers at least once every three weeks. The provider receives 70% of its funding in the form of a monthly service fee; the remaining 30% is contingent on achieving set targets. One third of the conditional payment is for successfully enrolling offenders on the programme. Another third is for achieving ‘hard outcomes’, such as clients entering employment or enrolling in further learning. The remainder is for meeting what the programme refers to as ‘soft outcomes’, but which are in reality a combination of outputs and processes, such as helping clients open bank accounts, organising mentoring and providing in-work support.

In early 2010, the Social Market Foundation proposed a model in which ten regional prime providers would operate small prisons dedicated to the incarceration of short-term prisoners. Providers would manage offenders both in custody and the community. They would receive a payment for the secure and humane incarceration of prisoners and outcome payments every six months for two years after a cohort of offenders had been released, based on the number who had not been reconvicted during each six-month period.92

Finally, in December 2010, the government announced it would be commissioning six pilots to test payment-by-results in offender management. Two of these will be aimed at offenders on community sentences and those released on
licence while another two will target offenders sentenced to less than 12 months in custody. It seems likely that at least one of these pilots will be jointly commissioned and will focus on an output such as drug use cessation or employment as well as reduced re-offending. A further two pilots will involve local partners working together to reduce re-offending. They will be able to retain a share of any savings made, to ‘be reinvested in further crime prevention activity at the local level.’

### Alternative Service Models

Even where providers are encouraged to explore alternative service models with the intention of accelerating the rate of reduction in re-offending, commissioners will have their own views about the linkages between outputs and outcomes, for the purposes of regulation and certification as well as the more effective negotiation of contracts and the management of providers.

Compared to services such as healthcare, where some linkages between drugs and clinical outcomes are well established, there is a lack of widespread agreement about what works to reduce re-offending. While evaluations of rehabilitation programmes no longer conclude that ‘nothing works’, one well-respected criminologist has observed that the evidence now supports the conclusion that ‘some things work for some people, some of the time, in some settings’. This lack of agreement about what works underlines the potential benefits that might arise from stimulating greater innovation, but it also helps to explain why commissioners cannot just specify high-level outcomes and allow delivery models to remain a ‘black box’.

Two different literatures inform the debate about how best to reduce recidivism. One is theoretical, seeking to understand why offenders desist. The contributors to this literature are psychologists, sociologists and criminologists who test their theories using longitudinal studies. Practitioners can draw on these theories and the accompanying evidence to design interventions to speed the process of desistance. The other body of evidence has been called the ‘what works’ literature. Rather than theorising the reasons for desistance, these authors evaluate programmes to identify those that have been most effective.

There is some convergence in the findings of these two literatures. Subscribing to one theory does not necessarily lead to the exclusion of certain types of programmes, although it is likely to lead to some being given more weight. As one source has described it, all programmes carry within them ‘implicit criminologies’ – assumptions
about why offenders start and cease committing crime and understandings of how programmes aid the process of desistance. The commissioner’s theoretical framework will have a powerful influence on the design of any payment-by-outcome system, and the specification of a particular measurement regime may lock out alternative approaches to rehabilitation, thereby narrowing the scope for innovation.

Resettlement approaches: Voluntary organisations have provided services to prisoners leaving custody since the 19th century, but in England and Wales, resettlement gained new importance with the publication of the Social Exclusion Unit report in 2002, which argued that prisoners’ needs upon release were not being met and that this was contributing to the high rate of recidivism. The government’s response was the National Action Plan, which identified seven distinct pathways, six of which were based on public services, while only one focused on ‘attitudes, thinking and behaviour’. The emphasis was on services that helped offenders manage daily life in the community, and the continuity of those services across the boundary from prison to the community.

It has been argued that ‘through the gate’ interventions of this kind are, at their root, determinist in nature reflecting a view that ‘offenders are largely the victims of their social circumstances and problems beyond their control’. Even if this description is unfair, it is true that characterising the problem in this way has resulted in a significant emphasis on structural and managerial reforms.

Motivational approaches: There is a multitude of competing theories that can be categorised as motivational, but they differ from resettlement theory in their focus on individual agency, rather than structure, in the process of desistance. Change is seen as ‘a difficult and often lengthy process’, with numerous relapses and reversals. Thus programmes based on motivational theory tend to place much greater focus on the development of human capital.

Motivational approaches to desistance can live harmoniously alongside a resettlement approach, recognising the importance of helping offenders to cope with the practicalities of daily life. Indeed, it has been argued that the two approaches are mutually reinforcing; solving the practical problems without addressing thinking and attitudes or vice versa is unlikely to be effective.

The Ministry of Justice recognises the importance of a mixed approach. The 2010 Green Paper Evidence Report cites ‘good evidence that cognitive/motivational programmes… can reduce re-offending; and there is promising evidence about the impact of drug treatment programmes [and] education, training and employment’.
Moreover, the report emphasises the importance of the supervisory relationship between offender and case manager to rehabilitation and reduced re-offending.102

However, the foundation of NOMS’s approach to desistance has been the more effective coordination of services ‘through the gate’: ‘the [regional] plans outlined under most of the Pathways seem to take it for granted that good service provision will result in less re-offending…’ Given limited time and scarce resources, there is an even greater danger that ambitious plans for one-to-one supervision of offenders, as foreseen in the Green Paper, may be compromised by more instrumental approaches.103

Selecting Measures

**Accurately Measuring Outcomes**

Implementing payment-by-outcome requires commissioners to choose outcomes and one or more measures to assess and reward attainment. While criminal justice programmes often measure reductions in reconviction rates, interventions can also target outputs such as increasing employment rates.

The primary outcome sought by criminal justice programmes is desistance from crime. Theoretically, this could be measured by a reduction in the rate of re-offending; however, it is impossible in practice to measure this. Indeed, it is extremely difficult to arrive at reliable statistics on the rates of crime overall, which means that many individuals classed as first time offenders by the system will in fact be re-offenders. Moreover, even if all crimes were detected, it would still be necessary to attribute them to known offenders in order to measure rates of re-offending.

As a result, payment-by-outcome schemes must rely on intermediate outcomes which are broadly indicative of the level of re-offending and can be measured, such as reductions in the rates of reconviction or re-imprisonment. Alternatively, commissioners might specify a cluster of outputs that they believe contribute to the reduction of re-offending. Reductions in drug misuse, improvements in the stability of relationships, and success in becoming debt-free, although not themselves indicative of desistance, are considered to be important factors in reducing re-offending. Intermediate outcomes are likely to be more reliable measures as they are more strongly correlated with reductions in re-offending.

In selecting outputs or intermediate outcomes to measure, commissioners must be satisfied that they are reliable proxies for the primary outcome. For example,
the length of time over which outcomes or outputs are measured will impact on their utility as proxies. Of offenders who completed sentences between January and March 2000, 43% were reconvicted within one year of their release; by the end of nine years, this figure was 75%. Clearly, measures of non-reconviction taken at the end of nine years would more accurately reflect actual desistance than measures taken at the end of the first year. However, longer measurement periods increase transaction costs, so that increased reliability of proxies must be weighed against monitoring costs. Ministry of Justice statistics show that of offenders who completed sentences between January and March 2000, more than three-quarters of those who re-offended in the two-year follow up period did so within the first year of measurement. This suggests that the advantages of a longer measurement period decline substantially after the first year.

Clear, Assessable and Continuous Measures

Ideally, proxies should also be clear and easy to measure. They should also be continuous; that is, where commissioners value progress towards an outcome, measures should reflect such progress. However, it is very difficult to find a measure that has all of these characteristics.

There are two main ways of measuring reconviction rates. The first is known as a binary measure, and simply records whether or not, in a certain period, offenders have been reconvicted. The second is referred to as a ‘distance-travelled’ measure and captures more qualitative elements of reconviction, such as how many reconvictions have occurred over a defined period of time and how severe the offences were. Both approaches have advantages and shortcomings that commissioners will need to weigh when making their selection.

In offender management, binary measures assess absolute desistance by assessing whether or not offenders have been reconvicted. One advantage of such measures is that they give providers a clear indication of what they must achieve: they send a strong signal that anything less than complete desistance is a failure.

Binary measures are more amenable to statistical analysis. In order to measure a provider’s impact on outcomes, commissioners must be able to compare their results with those of a control group or a statistically-derived level of expected performance. The use of a control group is considered to be more rigorous. Where for logistical reasons this is not possible, for example where all offenders are receiving post-release services, commissioners will need to compare actual
outcomes to statistically-predicted results. The Ministry of Justice generates predictions of the percentage of offenders released in a particular year who will re-offend within 12 months of their release.

As binary measures of offending simply inform commissioners whether or not offenders have been reconvicted, they provide limited information about changes in offending behaviour. Distance-travelled measures such as reductions in the frequency or severity of offences, allow commissioners to assess the progress providers have made towards achieving outcomes. Criminologists argue that offenders do not abruptly desist from crime but rather gradually reduce the frequency and severity of their offending.107 Given the nature of desistance, it may be considered unfair to penalise providers who had achieved reduced frequency and severity of offences where some of their offenders were nevertheless reconvicted.

One advantage of distance-travelled measures is that they may incentivise providers to engage with high-risk offenders who are unlikely to achieve absolute desistance and who providers may not otherwise wish to serve.108 On the other hand, distance-travelled measures may require more complex measurement systems. For example, offenders who are re-incarcerated are unable to commit any further recorded offences, which may create challenges for commissioners attempting to measure frequency. Commissioners could solve this problem by ‘pausing’ the monitoring of offenders while they were imprisoned and restarting once they were released, but this could make tracking a cohort complicated as there would be multiple programme termination dates. While there may be solutions to this problem, this does suggest that using distance-travelled measures results in much more complex measurement systems.

Finally, it is more difficult to attribute the impact of providers with distance-travelled measures. Because there are currently no predictions of the frequency or severity of re-offending due to the complexity of constructing such measures, control groups would be needed to attribute providers’ impact on distance-travelled measures.109

Controlled Innovation
One of the advantages of paying for outcomes rather than processes is that providers can experiment with different ways of delivering programme objectives. Commissioning for a reduction in re-offending would allow providers to test alternative service models, including resettlement and motivational approaches,
in the search for a more successful way of reducing re-offending. However, as it is not possible to observe the primary outcome directly, commissioners must measure intermediate outcomes, such as reductions in the rate of reconviction or re-imprisonment of a cohort, or outputs, such as success in finding stable employment and accommodation, in order to measure achievement.

Where commissioners use intermediate outcomes as measures, the scope for innovation by providers is still likely to be considerable since providers may take quite different approaches to reducing reconviction or re-imprisonment rates. Where commissioners decide to measure and reward outputs, however, innovation is likely to be constrained as some outputs will not be compatible with some alternative service models.

For example, Job Deal is clearly based on resettlement theory as finding employment is part of coping with daily life outside of prison. Providers of this programme are not free to experiment with motivational approaches since they have been contracted to provide employment-related services, and part of their payment depends on their achieving ‘soft’ and ‘hard’ employment and education targets.

Commissioners may deliberately choose to limit the extent to which providers are able to operate programmes based on very different service models, either because they consider the evidence for those approaches to be weak or because the programmes would be politically unacceptable. For example, paying offenders to desist could be controversial.

Thus, different kinds of measures can be used to control the level of innovation by providers. Commissioners should take care to use this tool consciously, ensuring they are fully aware of how using a particular measure could narrow or broaden the scope for research and development.
Case Study 3: Long-Term Condition Management

Background
Long-term conditions, the most common of which worldwide are heart disease, stroke, diabetes, asthma, cancer and chronic obstructive pulmonary disease, are diseases that cannot currently be cured but can be controlled with the use of medication and/or other therapies. As of January 2010, 15.4 million people in England were living with a long-term condition and the number of people with at least one such condition is expected to rise to 18 million by 2025. The costs associated with treating these conditions are large, accounting for 70% of the total health and social care spend in England. By 2022 public expenditure on long-term care is expected to rise by 94% to £15.9 billion, and given the growing prevalence of these conditions and the concomitant rise in healthcare costs, there is a strong case for exploring new ways of ensuring that they are systematically and proactively managed.

Most chronic care programmes in the United States, Australia and the United Kingdom are evaluated in part based on the outcomes they achieve. UnitedHealthcare’s Evercare programme, for instance, defines programme success by the number of hospital admissions avoided through shifting care for frail elderly patients from the hospital to the nursing home. In the United Kingdom, the Evercare model has been adapted to work in a community setting. In the Newham pilot, patients at high risk of suffering another hospital admission in the 12 months following discharge were assigned a community matron who was responsible for ensuring that the patient did not relapse. In East Lincolnshire, an integrated approach to managing chronic obstructive pulmonary disease (COPD) was taken,
and a specialised COPD intermediate care team called ‘Inspire’ was established, spanning primary and secondary care. Outcomes measured included reductions of admissions, re-admission rates and length of stay, improvements in quality of life indicators and mortality data.

However, while outcomes of chronic care programmes have been monitored and published, providers have so far not been paid by outcomes. Instead, physicians in the US, UK and Australia are being paid for outputs and processes thought to improve quality of care.

In Australia, GPs have received blended payments since 1999 in an effort to move away from a fee-for-service model. In 2003, outcome payments were introduced, providing additional remuneration for doctors completing certain treatments or tests for a percentage of the population in each disease area. With diabetes, for example, the main indicator of quality is whether a test of blood sugar levels is conducted during the consultation.111

In the United Kingdom, payment-by-results was built into GPs’ General Medical Services (GMS) contract in 2004 in the form of the Quality and Outcome Framework (QOF). This offers financial rewards to practices that demonstrate their achievement on 128 quality indicators in four domains. Points for clinical quality are awarded where practices can demonstrate that they have fulfilled a number of key stages in the management of chronic disease for a proportion of the relevant population.

Since 2001 hospitals in the UK have also been placed under a payment-by-results regime. All providers of hospital care are paid nationally-determined fees based on the number of predefined activities (called Healthcare Related Groups) carried out.112

In 2000 the US Congress mandated the Centers for Medicare & Medicaid Services (CMS) to test a hybrid payment methodology for physician groups that combines Medicare fee-for-service payments with incentive payments. The participants were eligible to earn annual incentive payments by achieving cost savings and meeting quality of care targets. In the first year, physicians were assessed against six quality targets set by the CMS, such as whether a beneficiary’s blood pressure was at the recommended level. In 2001, the California Pay-for-Performance Program became the largest non-governmental physician incentive program in the United States. Performance is assessed based on clinical process measures, and, since 2005 physicians are also remunerated for achieving targets, such as reducing diabetes patients’ blood sugar levels to a certain threshold.
The insights derived from the outcome evaluations of long-term condition management and result-based payment schemes for physicians, such as the Quality and Outcomes Framework, can be used to analyse how payment-by-outcome might apply in long-term condition management and the issues that might arise in the process. The next sections examine why payment-by-outcome is desirable and what tools are available to improve the chances of success.

**Why Pay for Outcomes?**

Payment-by-outcome offers two particular advantages in the field of long-term condition management. It fosters innovation by giving providers the flexibility to experiment with the linkages between inputs and outputs, and outputs and outcomes, and it may improve efficiency if providers are allowed to retain cost savings associated with treating patients proactively.

Of course, much innovation in the medical treatment of long-term conditions has already occurred. The scientific evidence on the effectiveness of treatments is strong and the National Institute for Health and Clinical Excellence (NICE) already issues best practice guidelines for diagnosing and treating certain conditions. For example, there is compelling data on the role of exercise, diet, blood sugar control (HbA1c) and insulin injections in diabetes management. Providers are therefore unlikely to deviate from these treatments, especially since new ones have to be approved by NICE if they are to be funded on the NHS.

However payment-by-outcome may foster two other forms of innovation. First, providers may experiment with different ways of encouraging patients to cooperate with treatments. Often the effectiveness of treatment depends not only on the biochemical effects on patients’ bodies, but also on whether patients cooperate in the management of their diseases. In 2005, the UK government recognised the importance of patient involvement in their own care when it pledged to triple investment in the Expert Patients Programme, which delivers ‘free courses aimed at helping people who are living with a long-term health condition manage their condition better on a daily basis.’\(^{113}\) In the United States, healthcare organisations contact patients when they are more likely to be receptive to medical advice, such as when they receive a new diagnosis, experience changes in medication or are discharged from hospital.\(^{114}\) The timing of advice and treatments is thought to have a significant impact on patients’ willingness to co-produce.
Second, payment-by-outcome is likely to encourage providers to create functional service delivery systems. While there is scientific evidence of which treatments work for certain conditions, these treatments are not always administered to all patients who need them. Where providers take on the risk for health outcomes, they are incentivised to ensure that best practice is systematically carried out. This could be through standardised processes that ensure health indicators are monitored and recorded and treatments are changed or emergency care is given where needed, or through information technology that provides early warning of patient deterioration. Providers are likely to invest in finding new ways to ensure all patients receive the best possible care, as this will contribute to the achievement of outcomes.

In addition, payment-by-outcome may encourage providers to become more efficient, if providers are allowed to retain a share of the cost savings they produce. There is evidence that paying doctors for performing certain procedures, as is the case under payment-by-results, ‘encourages resource consumption’. 115 Although not yet tested, the theory would suggest that paying physicians based on patient health outcomes and healthcare utilisation, and allowing them to retain a proportion of the cost savings generated, would encourage them to be more cost-conscious.

Tools
Paying physicians by outcomes has clear benefits but is also challenging. This section analyses some of the tools available to commissioners that may make payment-by-outcome work successfully in long-term condition management.

Measures
The existence of measures that simultaneously proxy for outcomes and cost savings facilitates the introduction of payment-by-outcome since savings can be used to remunerate high-performing providers. In long-term condition management, a reduction in hospital utilisation – a function of frequency of hospitalisation and length of stay – has a direct impact on the cost associated with treating a particular patient and also proxies for health outcomes. Managed care programmes in Australia, the UK and the US are already evaluated on the basis of reduced patient hospital utilisation.

However, reduced patient hospital utilisation may not be a suitable measure for all types of patients. While a physician’s success with patients who have a prior history of hospitalisation can be judged by reductions in utilisation, the same cannot
be said for patients with mild conditions. Such patients often have no prior history of hospital episodes and may not be at imminent risk of admission. Reductions in GP visits or nurse consultations are alternative outcome measures for lower-risk patients. The number of GP visits is already factored into patient risk assessment tools such as the Combined Predictive Model used to predict patients’ risk of hospitalisation and it is also being tracked by the UK Department of Health as part of the ‘Whole Systems Demonstrators’ evaluation of telecare and telehealth pilots.

Furthermore, reduced patient hospital utilisation or service use need not be the only intermediate outcome that commissioners measure. In fact, unitary performance measures often do not adequately capture all aspects of primary outcomes and may induce providers to game the performance measure by improving measured performance without delivering better health. Therefore contracting for improvements in a number of outcome measures such as mortality and morbidity rates and service quality in addition to reductions in service use may be preferable. Monitoring a number of indicators gives the commissioner a more accurate picture of a provider’s contribution to the primary outcome and should reduce the incidence of creaming and parking.116

The Quality and Outcomes Framework, for instance, rewards GPs for achieving a number of clinical and organisational targets as well as for providing a good treatment experience to patients. Treatment experience is assessed using standardised surveys that ask patients a set of questions to reveal how they feel about their treatment. In 2009, Patient Reported Outcome Measures (PROMs) were introduced into the UK National Health Service (NHS). These are outcome- rather than process-based measures and are derived in two stages. First, patients are asked to report on various dimensions of their health (for example, mobility and pain/discomfort) in order to compute their overall health state. A preference weight or utility is then attached to that state by asking patients or a sample of the general population how many years of life in the current health state they would be willing to give up for a year of better health.117 Insofar as the trade-off forces patients to measure quality of life in a common currency, namely in terms of years of life they are willing to give up for better health, the scores of different patients can be compared. Theoretically, this means that commissioners could compare the scores of providers’ patients to evaluate their performance.

However, a major problem with both Patient Reported Experience and Outcome Measures is that providers may not have much control over key variables
that determine how much patients value a treatment or outcome. For instance, two patients both undergoing the same arthritis treatment could report different experiences as one is more sensitive to pain than another. Furthermore, two patients with the same mobility after treatment could report different PROM scores because mobility is more important for one patient’s job.

Awarding outcome payments to providers for achieving improvements in patient experience or patient reported outcomes may be unreasonable, since variables outside of providers’ control could influence the measures. Instead, monitoring and publishing these intermediate outcomes may act as a softer incentive for physicians to improve patient experience and quality of life.

In summary, there are a number of measures available to assess providers’ performance on primary outcomes in chronic disease management. While measuring patient hospital utilisation has the advantage of revealing information about a patient’s healthcare costs, it may be desirable to monitor other measures as well in order to better capture all aspects of primary outcomes and deter gaming.

Ownership- and Integration-Related Incentives
Payment-by-outcome is one means of aligning the interests of providers with those of the commissioner through a contract, but the termination of contracts may generate perverse incentives. When contracting a provider to manage a high-risk population nearing the ends of their lives, this may not be a problem. However, as commissioners seek to serve lower risk patients who may have 20 to 30 years left to live, the perverse incentives associated with a contract terminating at the end of, say, ten years, could be large.

These perverse incentives are best explained through the concept of ownership: who owns the benefit of a particular intervention. For example, if a provider is given a contract to manage the care of a population for ten years, and at the end of that period the contract is to be retendered, the subsequent provider will own the benefit of inheriting a population that has better health than at the beginning of the first contract. This means the original provider will have an incentive to innovate and manage the health of its patients proactively at the beginning of the contract period, because it is likely to benefit, but these incentives will dwindle towards the end of the contract. To ensure the original provider always has an incentive to manage patient health in the most long-term cost-effective way, the original provider must own the benefit of the intervention.
The integrated Managed Care Organisation (MCO) model in the US, which combines indemnity insurance with the provision of managed care to control healthcare utilisation and therefore costs, is one way of aligning incentives without the perversities associated with contract termination in a payment-by-outcome system.  

Most Americans are now enrolled in an MCO. Patients who are unsatisfied with the care they are receiving may be able to switch to another health plan, so the level of competition among health insurance companies is relatively high. Patients who are satisfied with the service, however, are likely to remain with their insurer throughout their lives. Competition gives MCOs an incentive to deliver very good quality of care, while the possibility of a client remaining with the insurer indefinitely ensures it owns the benefit of delivering cost-effective care over the long-term.

If doctors working on the frontline are self-employed or work for a number of different insurance companies, then they will not have the same incentives as the insurers to deliver proactive care. To better align the incentives of physicians, a number of MCOs operate as integrated care organisations, in which the insurer, commissioners and providers of care are organisationally or contractually integrated.

Kaiser Permanente, for example, has a largely contractually integrated structure. It contracts with the Permanente Medical Group of physicians on an exclusive basis, and, depending on the state, either owns and runs hospitals or contracts with non-Kaiser hospitals with which it has a long-term relationship. In this structure, Kaiser Permanente acts as both the insurer and the commissioner of care and physicians act as both providers and commissioners of secondary and tertiary care. This structure means commissioners can exert more influence over doctors to deliver outcomes, since physicians who contract with more than one insurer have the freedom to stop doing business with one if they choose, whereas doctors contracting exclusively with one MCO must either adopt the model of care of the organisation or leave their jobs. The high level of contractual integration means that Kaiser physicians ‘share a common destiny’ with the organisation as a whole and are therefore more likely to act in ways that increase its competitiveness, including delivering outcomes that improve the insurer’s financial position overall.

Furthermore, Kaiser doctors have financial incentives to ensure the company performs well. The doctors own shares of the MCO which ties part of their remuneration to its performance. As a result, physicians have a financial motivation to generate cost savings that improve the MCO’s profitability, for example
through managing long-term conditions in primary care settings. In fact, Kaiser Permanente performs 3.5 times better than the NHS in terms of the number of bed days it uses to treat 11 medical conditions for those aged 65 and above.

While the American system has been a reference point for some British policymakers, it is important to recognise that the UK system is fundamentally different since there is a purchaser-provider split. When considering how incentives will operate, this aspect of the UK context needs to be taken into account.

**Professional Norms and Regulation**

Payment-by-outcome creates a financial incentive for providers of long-term condition management to improve or prevent the deterioration of patient health. Depending on the design of the incentive system used to reward provider performance, providers may be inclined to game the system and increase measured performance without actually improving patient health. However, commissioners can harness professional norms to counteract gaming.

In the UK, healthcare providers are not only accountable for the quality of the services they deliver, as monitored by the Care Quality Commission, but they are also personally answerable for their professional conduct to the General Medical Council. The Council enforces professional standards as published in its general guidance, case studies and ethical standards. These clarify the meanings of good medical practice expected of physicians. Where physicians do not adhere to these professional norms, the Council has the legal mandate to sanction doctors and, in cases of severe misconduct, revoke their licence to practice.

Professional norms and regulation reduce the likelihood of gaming, as achieving measures without delivering outcomes violates professional standards to which physicians have legally committed themselves. For instance, under GP fundholding, doctors had a financial incentive to restrict access to secondary care (hospitals and specialists), as a reduction in referrals to secondary care promised to generate budget savings which doctors could reinvest in services in subsequent years. While such a financial incentive may have been expected to encourage doctors to postpone referring patients to secondary care even when they needed the treatment, this did not occur. Insofar as they are personally liable for negligence and misconduct, doctors are unlikely not to refer patients if it is in their best interests to receive secondary care.
Pharmaceutical Pricing

In the health sector payment-by-outcome has been used most extensively for pharmaceutical products. As governments and insurers have sought to contain rising health expenditures through greater scrutiny of cost-effectiveness, drug manufacturers have assumed some of the risks around the performance of new products.

The transfer of outcome risks was first developed in the US in the 1990s between insurers and manufacturers. Early examples included ‘no cure, no pay’ strategies for male pattern baldness drugs, schizophrenia treatments and cholesterol-lowering statins. This approach proved unsustainable as it became clear that the schemes benefited insurers far more than manufacturers, and they were either not renewed or manufacturers attempted to renegotiate the agreements, which generated mistrust.

In 1999 an outcomes guarantee was piloted in North Staffordshire (UK), with the trial of a new branded statin. The manufacturer promised to refund all costs of its drug if it failed to reduce patients’ LDL cholesterol to safe levels. This proved successful and allowed the makers to differentiate their product from older statins that were due to go off-patent and decrease markedly in price.

Recent outcome-based schemes in the UK have been applied to more complex products. The first national scheme arose amidst controversy following a NICE decision not to recommend drugs which slowed the progression of Multiple Sclerosis. As a result, in 2002, the Department of Health negotiated a scheme whereby list prices would be reviewed so as to meet a maximum cost-effectiveness ratio of £36,000 per Quality Adjusted Life Year (QALY, a measure of the quantity and quality of life generated by a healthcare intervention). In Australia a similar
scheme was developed to adjust the price of a drug for a rare pulmonary disease based on mortality rates.\textsuperscript{130}

Simpler rebate schemes have been used for two cancer drugs, Velcade and Erbitux. The manufacturers offered to refund costs after a pre-agreed course of treatment if the outcomes fell below the expected levels.\textsuperscript{131} Another drug company offered to bear the cost of treatment for patients with macular degeneration after 14 doses of their Lucentis product if there was no improvement in visual acuity compared to standard care.\textsuperscript{132}

In the US, insurers have turned once again to risk-sharing schemes. The manufacturers of an anti-osteoporosis drug have reached an agreement with a medium-sized insurer to reimburse costs associated with fractures for patients taking the treatment as prescribed. The manufacturer Merck agreed with a major insurer, Cigna, to discount the cost of its anti-diabetes drugs following decreases in blood-sugar levels of patients taking \textit{any} diabetes drug and further discounts if patients were taking Merck drugs as prescribed.

There have also been movements to extend value-based purchasing to all branded drugs. The UK Pharmaceutical Pricing and Regulation Scheme was revised in 2009 to reflect this concept, and a pay-for-performance model has been proposed by the Centers for Medicaid & Medicare Services.\textsuperscript{133}

\textbf{Outcomes and Measures}

Research and development of novel drugs is rooted in a scientific process of discovery and measurement which gives purchasers and manufacturers a certain amount of confidence to implement payment-by-outcome schemes. However, much of the evidence on pharmaceutical performance is produced under clinically controlled trial conditions, which addresses the drug’s efficacy in targeting a particular process whilst what matters most to purchasers is its effectiveness with real-world patients.

\textbf{Surrogates}

The performance of some products is more easily measurable where a clear biomarker exists. This is an objective and measurable biochemical feature that can be used as a surrogate for hard outcomes. Widely used biomarkers such as LDL cholesterol for heart disease or glycated haemoglobin for diabetes are based on well-evidenced links between use of the drug and the surrogate and between the
surrogate and disease progression. A clear example of this is the Velcade Response Scheme for patients with multiple myeloma (a cancer of the blood) where response was measured by a reduction of abnormal cells in the blood known as M-proteins. Nevertheless, there can be uncertainty around interpreting biomarkers. There are concerns that M-protein is not a good surrogate for life expectancy, whilst in 10-15% of cases patients do not have measurable M-protein levels.

### Functional Measures

Where clear evidence-based surrogates do not exist, functional measures based on physical improvement or harm have been used in fairly simple rebate/discount agreements. The Lucentis scheme for patients with macular degeneration passes on the costs of treatment to the manufacturer if there has been no improvement in eyesight compared to standard care based on visual acuity scores. Alternatively, the Actonel Fracture Protection Programme reimburses insurers for the costs of non-spinal fractures based on average medical expenses. The Erbitux scheme for colorectal cancer addresses the size of the tumour as interpreted by doctors.

### Effectiveness

More sophisticated measurement systems have tried to scrutinise more intensely the effectiveness of pharmaceutical products in the real world rather than simply their efficacy in altering biochemical features. An important feature of these schemes is that they measure the deviation of actual from expected performance rather than from a placebo or a control group and therefore assess whether the product lives up to the claims made based on clinical trial data.

In Australia a registry for patients diagnosed with pulmonary arterial hypertension and taking the drug bosentan was used to compare actual annual mortality rates to a benchmark rate derived from a predictive model based on evidence from clinical trials. An increase in mortality rates would lead to a decrease in price.

The UK MS risk sharing scheme uses a ten-year study to monitor disease progression in patients compared to an expected progression derived from a predictive model. Disease progression was measured using the expanded disability status scale (EDSS), specifically developed to measure MS patient health on a scale from zero (perfect health) to ten (death). Although the study began in 2002, it has not yet provided conclusive evidence concerning the actual effect of the drug. This is largely due to the uncertainty of the outcome measure which does not fully
capture the complexity and long-run nature of the disease and suffers from variation and measurement error.\textsuperscript{134}

**Cost-Effectiveness**
Incremental cost-effectiveness ratios based on quality-adjusted life-years are a useful measure for capturing efficiency and are commonly used when negotiating prices. However, establishing thresholds can be highly sensitive since it requires a maximum price to be put on patient health and life expectancy. NICE has established a notional upper limit of £20-30,000 per QALY, above which a drug will generally not be recommended. The Australian Pharmaceutical Benefits Scheme has set a similar ceiling at A$60,000. In the US, insurers have reported that state regulations and market pressures make it virtually impossible for them to refuse a drug.\textsuperscript{135}

Establishing cost-effectiveness is heavily dependent on the quality of data available regarding quality of life and life expectancy. Initial calculations of the cost-effectiveness of beta-interferons for MS based on randomised control trials prior to the risk-sharing scheme produced wildly divergent results ranging from £20,000 to £1 million per QALY.\textsuperscript{136} The interim results of the risk-sharing scheme’s 10-year monitoring study have failed to provide any further clarity.

**Financial Incentives**
Governments and health insurers must strike a balance between asserting control over high pharmaceutical costs and continuing to foster innovation in new products. Voluntary agreements with industry to regulate gross profits, such as the UK’s Pharmaceutical Price Regulation Scheme, and taxation aimed specifically at manufacturers, as used in the US, form a background to pricing incentives but are inevitably imprecise in their application.

The aim of outcome-based pricing schemes is to link manufacturers’ remuneration more closely to the performance of their pharmaceutical products and the actual value to the patient rather than covering the costs of research and development, marketing and manufacturing. There are three identifiable models:

a. rebate based on price of drug
b. rebate based on costs of harm
c. price adjustment (up or down) based on observed outcomes
In practice, prices tend to start high and gradually decrease over time. Increases are rare and usually arise in the renegotiation of prices as a result of higher manufacturing costs. Price-adjustment schemes usually only adjust downwards, thus manufacturers can mitigate their revenue risk if they can justify a sufficiently high entry price to cover input costs.

Rebate schemes can be high-stakes agreements since they depend on an all-or-nothing evaluation of patient response to trigger the refund. This makes the level at which the threshold response is set very important. The Velcade Response Scheme involved hard bargaining around the level of response required (reduction in serum-M protein biomarkers in the blood). Whilst NICE managed to negotiate higher standards of performance, raising the minimum response from 25% to 50%, patients who fell below this mark could lose out on further treatment despite experiencing some benefit. To mitigate this risk an extra cycle of treatment was allowed. However, given each cycle of treatment costs around £3000 this also increases the potential refund from the manufacturer.137

The burden of making claims tends to fall on the purchaser (in this case hospital pharmacy departments) and evidence from NHS patient access schemes for oncology drugs (such as Velcade and Erbitux) suggests that administrative costs are high and a significant proportion of rebates are lost as they are not claimed within the necessary time limits.138

A lower-stakes approach is used under the Lucentis dose-capping scheme where the manufacturer provides the treatment at no further cost after the first 14 rounds if an adequate response has not yet been achieved. This reduces the manufacturer’s potential exposure (since there is a guaranteed revenue stream), but caps the profit per patient. Moreover, it relieves the purchaser from some of the costs involved in assessing performance, since, in contrast to the all-or-nothing nature of the rebate schemes above, a claim only has to be made if further treatment is thought to be worthwhile.

Rebates related to the costs of harm can increase manufacturers’ exposure. The concept has so far been applied to only one US scheme for anti-osteoporosis tablets whereby the average medical expenses associated with a non-spinal osteoporotic fracture are reimbursed to the insurer. Depending on the type of the fracture this can cost up to $30,000 per patient compared to the cost of the drugs which is around $1000 per year. Successful claims are required to show that the patient has been taking the drug as prescribed for at least six months whilst a maximum
number of re-imbursable fractures was also set as part of the one-year agreement. Data from the first nine months showed that the reimbursement rate was well below the maximum for the year.\textsuperscript{139}

An alternative model of discounting prices based on improved performance has also been used by a US drug manufacturer. The agreement between the US insurer Cigna and Merck for the oral diabetes drugs, Januvia and Janumet, drives adherence and greater volume sales by providing one set of discounts if blood sugar levels of patients decrease and a second set if patients have been taking the drugs as prescribed. The insurer was chosen largely because they already offered their own diet and life-style programmes to help manage the condition, suggesting that manufacturers are willing to take greater risks with their money where they feel that purchasers and their patients will co-operate.

Manufacturers have also been willing to assume the costs of monitoring outcomes for the purposes of complex price-adjustment schemes. However, there is little evidence from the beta-interferon and bosentan price-adjustment schemes of changes in pricing so it is not yet possible to observe an effect on incentives.

**Identifying the Patient Population**

Pharmaceutical products are highly differentiated and specialised. They are therefore explicitly targeted at particular types of patients from a very early stage of development but even at launch there will be disagreement about who might benefit. Risk-sharing schemes have helped to overcome some of the tensions involved in price negotiations between manufacturers and purchasers by setting out clear inclusion criteria whilst also facilitating an ongoing refinement of the population as the scheme is rolled out and effectiveness is assessed.

**Inclusion Criteria**

Under randomised control trial conditions, manufacturers usually recruit a narrowly defined sample that is most likely to achieve optimal performance against a control group. Following successful clinical trials, however, manufacturers are keen for their treatments to be applied as widely as possible to increase profits, whilst purchasers will seek to narrow application to contain costs.

However, under a risk-sharing scheme the purchaser may be more inclined to open up the inclusion criteria whilst the manufacturer will be more careful to identify patients who are closer to the population that took part in the randomised
control trial. NICE decided to expand inclusion criteria for the age-related macular degeneration drug Lucentis, to include patients with deterioration in one eye after Novartis agreed to refund the cost of doses after the 14th injection.

Clinical and genetic markers are also used to predict the likelihood of response and therefore influence the selection of the appropriate population. Many new oncology drugs now have an accompanying test to identify the most likely responders. In 2007 UnitedHealthcare entered an 18-month risk-sharing trial with the maker of a US$3,500 genetic test which determined whether a woman with early-stage breast cancer would benefit from chemotherapy. The insurer paid for the test during the trial in the expectation that a lower price would be negotiated if the costs of chemotherapy had not reduced in line with test results.140

Continuation Criteria
Once risk-sharing schemes are in progress, it will become apparent that some patients fail to respond as expected and do not benefit from the drug, necessitating further refinement. Whilst a randomised control trial cannot exclude participants based on sub-optimal performance, in real-world conditions, physicians will use their judgement in line with professional guidelines when observing patients on a course of medication and decide whether to alter dosage, switch to different medication or stop treatment altogether. Targeting the most responsive patients also activates a virtuous circle of improvement since patients will be more likely to keep taking a drug as prescribed where they experience benefits.

Continuation criteria are designed to increase the cost-effectiveness of risk-sharing schemes and are generally stricter than original guidelines. In the Australian bosentan scheme, patients were rigorously assessed every six months and 20% of patients were excluded if their conditions did not stabilise or improve, since they were unlikely to benefit from continuing to take the drug.

Foster Care
Foster care is a short-term measure that is widely regarded as inadequate in meeting the physical, social and emotional needs of children over the medium- to long-term. Those children who do not find a permanent home encounter more problems in the future: they commit significantly more crime, spend more time in jail, and receive disproportionately high welfare assistance as adults.141
In the United States, the management of child welfare is divided: while states continue to have primary responsibility for provision, the federal government has intervened to impose standards and provide technical assistance. Not-for-profit organizations have assumed responsibility for actual delivery funded through grants and contracts. In some states, adoption and foster care services have been contracted together, since fostered children often find permanent homes with their foster families and both services require similar skills. In other cases, only adoption services are outsourced, so agencies arrange permanent homes for those already in foster care. Other states have created different markets for foster care and adoption. However, the identification and prevention of abuse in the original family environment remains a separate part of the service and is typically performed by a public sector agency.

A payment-by-outcome approach was employed both in agreements between the federal and state governments, and in the contracts states signed with voluntary sector agencies. Child and Family Services Reviews were introduced in the year 2000 as part of a federal performance budgeting initiative. Some states, including Illinois, Michigan and Kansas, also implemented payment-by-outcome in their contracts with agencies. The design of these schemes reveals important insights into how the difficulties associated with payment-by-outcome have been managed.

One of the principal challenges lay in selecting effective measures of performance. Child and Family Services Reviews monitor the performance of states according to safety, permanency, and child and family well-being. These categories are broken down into seven intermediate outcomes and 23 indicators such as the percentage of children re-entering foster care after being reunified with their families and the percentage of children placed with siblings. States that do not achieve the required outcomes may accrue financial penalties which can run into millions of dollars, although as of January 2004 no financial penalties had been applied. While these reviews are acknowledged as having instigated a revolution in performance management, there were problems with the measures used to assess performance.

It has been argued that the indicators ‘fail to capture experiences of children and families that adequately reflect safety, permanency and well-being outcomes.’ In addition, the methodology underpinning data collection has been criticised for skewing the measures. Measures are based either on a cross-section of children in
care or on those who exited care during a given review period. The former measure results in the over-representation of children with long lengths of stay, while the latter results in the ‘underestimation of length of time to permanency outcomes because of the bias of exit cohort samples towards children with shorter lengths of stay.’ If information were being used purely for improvement purposes, then it would still be useful, even if biased, as long as the methodology remained the same over time. However, when states face financial penalties for poor performance, accurate, unbiased data is essential.

Another important challenge of payment-by-outcome lies in determining the amount of the incentive payment providers should receive. Public sector agencies often do not have good quality data on costs, and setting a price that will motivate providers to perform is difficult. Kansas experienced problems with its first payment-by-outcome contract due to insufficient understanding of financial issues. Costs for foster care were 65% higher than estimated, creating fiscal problems for providers: the state paid foster care providers US$105 million in unexpected costs above the US$179 million contracted for, while the adoption provider received an additional US$31.4 million above the contract amount of US$37.4 million and yet was close to bankruptcy by the end of the four year contract.

Designing incentive systems that encourage appropriate behaviour can be difficult. When Kansas first outsourced its foster care and adoption services in 1997, the state paid providers a flat outcome payment (US$12,860-15,504 in 1997) per case. For this fee, providers were expected to deliver all traditional foster care services, and in addition, provide services for 12 months after a child was placed, with no additional funding if the child re-entered care during that time. Moreover, providers had to maintain standards in relation to child safety.

This incentive system was intended to encourage providers to place children quickly and deliver permanent and safe outcomes, but it was problematic since it did not pay providers more for existing foster care cases which were more difficult and costly to manage than new referrals. The second set of contracts let in 2000 changed the payment mechanism so that providers were remunerated based on a fixed price per child per month, which meant that providers earned more for managing cases which were more time-consuming. Despite fears that the payment structure would create a perverse incentive for providers to keep children in foster care longer than appropriate or necessary, there is still a positive incentive for providers since annual contract renewal depends on performance.
Michigan may have experienced similar problems. While the focus of foster care and adoption is to find children safe, permanent homes, the amount of time children spend moving among temporary placements also has an impact on their later lives. Since 1992, Michigan has incentivised agencies to find children good homes more quickly. Once a child is removed from her original home she is placed with the state or private agency with the best available foster care home at that time. This agency then has a six-month window to secure an adoption (proposed adoptive parents have to be found within three months to ensure a stable, well-planned transition). If after six months the child has not been placed, her details are made available to all licensed agencies which can then compete for placement. Agencies are paid primarily according to the speed with which adoptions are finalised.

Since providers are paid less for children not placed after one year than for newly-available children, concerns that providers will not be as strongly motivated to place such children may be legitimate. Available data neither confirms nor denies this: the total number of adoptions between 1991 and 1999 increased by 83%, but the number of children available for adoption increased 116% over this period, suggesting that providers did not manage to place even all the newly-available children; however, a breakdown of the proportion placed by length of time in foster care, which would confirm the hypothesis, is not available.

Finally, extraneous variables appear to be a significant issue in payment-by-outcome in foster care and adoption. Agencies rewarded for placing children in safe, permanent homes cannot control the number of children who require foster care or adoption (which depend on, among other variables, the success of preventive services and court decisions), and yet this is one of the main variables that will impact on their results. Because of this, contracts that shift small amounts of risk to providers may be more suitable in this sector than high-stakes contracts. For example, in Cook County, Illinois, agencies were paid a monthly fee per child and were expected to place 24% of children into permanent homes each year. Failing to do so would result in the state considering not referring any additional children to the agency in the future. In other parts of Illinois, providers received US$2000 for each child placed after the 24% rate has been achieved. This relatively soft form of incentivisation nevertheless achieved dramatic results. Adoptions rose by 94% from 1997 to 2003 and the permanency rate rose from 2-4% to 12-23%, while costs fell.
Foreign Aid

Significant change has taken place over the past three decades in the conditionality of international aid, with a shift from payment for promises, with grants and loans linked to commitments to policy change, to payment for processes, where recipient governments demonstrate their commitment by making institutional change aimed (say) at reducing corruption and increasing local involvement. Among other things, process conditionality has been criticised for being intrusive and undermining local accountability.

The difficulties with conditionality serve to highlight the fundamental problem that outsiders cannot be effective and can actually do harm if they try to design the ‘software’ of an economy.\textsuperscript{154}

Starting around the year 2000, the international community began to focus on performance-based funding, with the Paris Declaration in 2005 committing donors to increasing the ownership of initiatives by recipient countries, improving accountability and focusing on results and outcomes as measures of performance.\textsuperscript{155}

A number of foreign aid schemes in recent years have adopted payment-by-results. The Global Alliance for Vaccines and Immunization has made new grants conditional on countries’ past performance, with payments linked to the number of children vaccinated. In a programme sponsored by the UK Department for International Development, the World Bank has introduced ‘output-based aid’ in some of its water projects, with concessionaires paid for the number of water connections. Conditional cash transfers make payments dependent on recipients’ performance of some key activity, such as ensuring that children attend school or visit health clinics on a regular basis.\textsuperscript{156}

More recently, Nancy Birdsall, President of the Center for Global Development in Washington D.C., has proposed a comprehensive system of cash-on-delivery incentives in foreign aid, with particular application to completion rates in primary schooling: ‘The core of COD Aid is a contract for funders and recipients to agree on a mutually desired outcome and a fixed payment for each unit of confirmed progress.’\textsuperscript{157}

Streetscene Management

An integrated approach to improving the quality of public spaces has been a notable trend in local government commissioning. Local councils have sought
to bring together services such as waste management, recycling, street-cleaning and parks maintenance which enhance their residents’ shared experience of local areas. Councils have also begun to consult more actively with their local residents in order to understand preferences and local priorities. Some councils have gone further and negotiated agreements where a portion of the contractor's remuneration is linked to user satisfaction with the services as well as increased recycling rates, removing graffiti and tackling fly tipping.

In 2003 Woking Borough Council concluded a ten-year contract with a private provider for street-cleaning and landscaping services. The contract does not stipulate any inputs and the provider is free to adjust schedules and redeploy the workforce in the most effective manner. The provider is paid an annual service fee, which is supplemented by a performance payment of up to 8% of the total annual contract value which represents the provider's profit margin. The performance fee is calculated using a sliding scale of overall customer satisfaction with the cleanliness of streets and the appearance of parks, flowerbeds and grass verges measured with a quarterly telephone survey of 350 local residents and administered by an independent third party. The scale runs from 60-100% overall satisfaction, so any outcome below 60% elicits no extra payment beyond the service fee, whilst three continuous quarters of sub-threshold performance permit the council to terminate the contract. These performance incentives have caused the provider to experiment with ways to increase satisfaction with services. For example, cards are delivered to residents to alert them to the fact that their streets have been cleaned and seek feedback.

In Charnwood, the same provider of ‘street scene’ services has introduced a ‘Community Champions’ scheme which equips volunteers with digital cameras fitted with GPS devices to take photographs of local ‘grot spots’ blighted by fly-tipping, graffiti and littering and send them to a rapid response team which can very quickly arrive and cleanse the affected area.

In Sandwell, a deduction-based model has been adopted using a blended suite of indicators which includes customer satisfaction targets, recycling rates and environmental cleanliness. Deductions are accrued for every performance failure which are then subtracted from monthly payments to the provider. Failure to rectify performance failures can lead to a multiplier being used to motivate the provider to act rather than absorb the cost. To mitigate the downside risk for the provider, during the first year a quarterly allowance is in place which allows them to accrue
a certain number points without a deduction being made, similar to an insurance excess. However, after the 12-month ‘bedding in’ period, this allowance will no longer be available.
Conclusion

Since it was designed as a toolkit, this report contains a number of detailed insights into how payment-by-outcome can be made to work more effectively. However, three over-arching design principles have emerged as particularly important. Two of these – the warning that payment-by-outcome has its limitations and must be understood if it is to be used appropriately, and the need to build systems over time using the full range of tools in the toolkit – have already been discussed in some detail in body of the report, so it is with the third that this report concludes.

Where the linkages between outputs and outcomes are already well understood, and where they are tightly-coupled so that the successful completion of a process almost invariably leads to the desired outcome, then there is little purpose in attempting to transfer outcome risk from commissioner to provider. On the other hand, where there is very little understanding of or agreement about the linkages between effort and outcome, it will be virtually impossible to write an outcomes-based contract that effectively transfers performance risk.

Payment-by-outcome seems to work best in situations where commissioners are confronted by ‘known unknowns’. It is in these situations where commissioners are able to transfer performance risk, and there are social and economic benefits from so doing. These gains may come from the discovery of more effective linkages between outputs and outcomes, so that, for example, commissioners may wish to use payment-by-outcome to encourage providers to explore innovative drug rehabilitation techniques that significantly reduce recidivism rates.

Or these gains may come from the better implementation of what is already known. For example, it may already be widely agreed that a certain pharmaceutical will improve the quality of life of a particular group of patients, but still necessary
to seek out and identify that sub-set of the population for whom the drug will make the greatest difference, and to motivate those patients to take their medication as prescribed.

Much of the current interest in payment-by-outcome seems to relate to this second kind of innovation, with particular emphasis on: (i) identification of those beneficiaries for whom particular service models will work best; (ii) creation of effective management processes (for example, through joining up fragmented supply chains) enabling services to be tailored to different classes of beneficiary; and (iii) encouragement of much greater co-production on the part of beneficiaries.

It is no coincidence that the sectors policymakers have earmarked for the implementation of payment-by-outcome are those that bear these characteristics.
Acknowledgements

The 2020 Public Services Trust and the authors would like to thank the many people that participated in the preparation of this report, generously giving of their time and offering encouragement, information and insights.

Project sponsors
Local Partnerships, Partnerships UK, Serco Group plc

Commissioner project lead
Lord Geoffrey Filkin

Advisory board members
Lauren M. Cumming, Alastair Dick, Lord Geoffrey Filkin, Gary L. Sturgess

Foster care case study
Peter May, Online Editor, The Serco Institute

Long-term condition management case study
Miles Ayling, Director of Service Design, Commissioning and System Management, Department of Health
Conor Burke, CEO, Redbridge PCT
Tim Ellis, Programme Manager for Whole System Demonstrators, Department of Health
Peter Forrester, Director, Serco Consulting
Pam Garside, Fellow in Health Management, Judge Business School, University of Cambridge
Nick Goodwin, Senior Fellow, King’s Fund
Jeff James, CEO, Wiltshire PCT
John Myatt, Director, Serco Consulting
Offender management case study
John Biggin, Contract Director, HMP Doncaster
Luke Edwards, Head of Strategy and Change, Ministry of Justice
Elizabeth Fells, Head of Public Services Reform, Confederation of British Industry
Steve Hall, Contract Director, Business Development, Serco
Chris Harrison, Contract Manager, Serco Welfare to Work
Andy Homer, Assistant Director Operations Support, Serco Civil Government
Stephen Hornby, Senior Partnerships Manager, Serco FND Manchester
Wyn Jones, Contract Director, HMP Dovegate
Richard Judge, Finance Director, Serco Welfare to Work
Martin McClellan, Senior Manager, Offender Management, HMP Doncaster
Verena Menne, Researcher, SMF
Phil Oliver, Senior Assistant Director, Security and Operations,
    HMP and YOI Doncaster
Andrew Templeman, Director, Serco Consulting
Nigel Thacker, Market Development Director, Reliance
Tom Thackray, Policy Adviser, Confederation of British Industry
Trevor Williams, Assistant Director, HMP Dovegate

Streetscene management case study
Robin Davies, Marketing Director, Serco Local Government and Commercial

Welfare to work case study
Mike Hope, Delivery Directorate Senior Analyst, Department for Work and Pensions
Richard Johnson, Managing Director, Serco Welfare to Work
Khusbu Patel, Bid Manager, Serco Welfare to Work

Report preparation
Heidi Hauf, 2020 Public Services Trust; Jeanette Thompson, The Serco Institute
    (administrative support)
Peter May, The Serco Institute (proofreading)
SoapBox, www.soapboxcommunications.co.uk (design and printing)
Endnotes

16. Ibid., 171.
17. James Q. Wilson, op. cit., 133.
18. See Jeffrey L. Pressman and Aaron Wildavsky, op. cit.
25. Ibid.
30. Ibid., 32.
33. This was Professor Marilyn Strathern’s reformulation of Goodhart’s Law.
36. Ibid., 9.
39. Ibid., 50.
43. Burt S. Barnow and Jeffrey A. Smith, op. cit., 29.
64. Please note these numbers are used for illustrative purposes only.
67. Most commissioners do specify a certain number of hours per week that clients must be in work in order for providers to be able to claim the outcome payment.
Please note these numbers are used for illustrative purposes only.


Peter Saunders, op. cit., 16

Ibid., 18.

Mark Considine, op. cit., 51.

Peter Saunders, op. cit., 18.


Ibid.

Quoted in ibid.

Ibid.


A reconviction event occurs when an offender is reconvicted for any number of offences at a single court appearance in the 12 months following release. As the measure counts each court appearance it is tightly linked to cashable savings.


Ministry of Justice (2010), Breaking the Cycle, op. cit., 42.

Ibid., 43.


Mike Maguire and Peter Raynor, op. cit., 27.

Ibid., 21-22.

Ibid., 27.

Ibid., 24.

Ibid., 24-25.


Ibid., 59.

Mike Maguire and Peter Raynor, op. cit., 27-29.


The authors have drawn on insights from Nancy Birdsall and William D. Savedoff et. al., op. cit., 38, 46.

It is also possible to attribute outcomes using before-and-after measures, which compare offending behaviour before and after a sentence. However, this is considered the least scientifically rigorous method of assessing impact, so should be avoided.


Ministry of Justice (2010), Breaking the Cycle, op. cit., 45.

Ministry of Justice (2010), Re-offending of adults, op. cit., 45.


117. Alternatively, patients or the general population can be asked to rate their overall health state on a Visual Analogue Scale (VAS), with the end points labelled best imaginable health state and worst imaginable health state. VAS scores are of most value when looking at change within individuals, and are of less value for comparing across a group of individuals at one time point. For more details see: Dr. Mary Ellen Wewers and Nancy K. Lowe, ‘A critical review of visual analogue scales in the measurement of clinical phenomena,’ *Research in Nursing and Health* 4 (1990): 227-236.
119. Individuals who are insured through their employers are usually not able to switch, and those with pre-existing conditions may not be accepted onto new health plans.
122. Ibid., 22.
123. Arguably, US healthcare spending is growing as a percentage of GDP, but this trend does not imply an inability of HMOs to contain costs. The major drivers of higher spending are the decline of managed care and the growth of consumer-driven health plans. For more information see Ronald Lage, Deborah L. Aspling and Gert P. Westert, ‘Current and future developments in managed care in the United States and implications for Europe,’ *Health Research Policy and Systems* 3 (2005): 5.
125. The Medical Act of 1983 gives the GMC the authority to foster good medical practice and deal firmly and fairly with doctors whose fitness to practice is in doubt.
134. George C. Ebers, ‘Commentary: Outcome measures were flawed,’ *British Medical Journal* 2693 (2010).
141. Erwin A. Blackstone et al., ‘Privatizing adoption and foster care: Applying auction and market solutions,’ 
142. Children’s Bureau, ‘Section III – Narrative Assessment of Child and Family Outcomes,’ accessed online at 
143. United States General Accounting Office, Child and Family Services Reviews: Better Use of Data and 
Improved Guidance Could Enhance HSS’s Oversight of State Performance (Washington, D.C.: United States 
145. Albert R. Roberts and Kenneth R. Yeager, Evidence-Based Practice Manual: Research and Outcome 
147. Blackstone et al., op. cit., 1038.
148. Ibid.
149. Ibid.
150. Ibid., 1036.
151. Ibid., 1037.
152. Ibid., 1039.
153. Ibid., 1039-40.
156. Ibid., 36-37.
157. Ibid., 17.