Outcomes framework at PR19

Report for Thames Water

October 2016

Economic Insight

CONTENTS

1. Foreword by Dr Harry Bush

PR14 focused on the outcomes that customers want from their companies; on companies engaging with their customers to help determine what these are; and on companies taking ownership of their resulting business plans. It represented a marked shift from previous Ofwat price reviews. The changes it introduced were generally welcomed by companies and other stakeholders - and regarded as a significant advance on what had gone before. But, as in other sectors, there are always lessons to be learned from such change. There are loose ends and rough edges that need to be dealt with, as I experienced in my own tenure at the CAA. It took us some time to fine tune the initial 'constructive engagement' process we put in place to enable greater dialogue between airports and their airline customers.

Regulators can be enormously assisted in the process of refining their approach where debate is not refracted entirely through the prism of company self-interest, but where regulated companies are prepared to commission independent research and advice. Importantly, such advice can objectively expose the issues that regulators face and propose potential solutions. I was fortunate enough to contribute last year to the regulatory debate in the water sector, through an UKWIR sponsored report (co-authored with John Earwaker) on customer engagement, which sought to offer Ofwat pragmatic solutions to some of the issues that PR14 exposed. This new report, by Economic Insight, is in the same vein. It seeks to identify some of the critical issues that arise in relation to the regulation of customer outcomes and to propose solutions, exposing the choices and trade-offs that Ofwat faces.

The report starts from the premise that the changes introduced by Ofwat in PR14 represented a genuine advance on what had gone before, and that PR19 offers the opportunity to build upon success. But doing so means recognising some of the imperfections in the PR14 process and planning to avoid them this time round. Interestingly, some of the resulting themes echo those in my UKWIR report on customer engagement; in particular, the need for clarity from Ofwat at an early stage of the regulatory process and for consistency between different elements of the process and the resulting regulatory judgement.

This report particularly highlights the need for clarity from Ofwat on which performance commitments will be deemed 'comparative' and how far, if at all, company by company variation will be permitted. This, in part, is to avoid companies undertaking research and work that might not, in the end, be required. Similar issues arose in PR14, over which matters fell within the ambit of Customer Challenge Groups and which did not. The resulting confusion, and diversion of effort, can tend to cloud perceptions of an otherwise successful process.

The report also highlights the need both to take account of the inherent uncertainty in any benchmarking of outcomes performance commitments - and to ensure consistency between outcome and cost benchmarking. On the former, there can be a tendency to take the numbers generated by benchmarking at face value. There is a natural predilection to make the most of apparently hard comparative data. However, some caution is in order. The inevitable uncertainties around methodologies used, and the data points assessed, mean that cost and outcome variances revealed by benchmarking cannot necessarily be attributed to substantive differences in company performance. Measurement issues can also cause variance. So, while

striving to reduce it, regulators need to recognise and allow for the likely persistence of uncertainty around the causes of benchmarking variance in their processes and judgments. To not do so ultimately runs the risk of harming customers.

Turning to consistency, the sheer intensity of the work that falls on regulators during the price control, and the inevitable need for individuals and teams to specialise, can lead to a silo approach, whereby one part of the process is, despite best efforts, inadequately related to another. In principle, there should be consistency between the benchmarking of costs and outcomes, because one can clearly affect the other. Moreover, without join-up, companies would be wrongly resourced for the outcomes they are tasked to deliver, with adverse effects on customers. That is: companies either struggle to deliver the outcomes which they have been asked to deliver or, conversely, are over-rewarded for doing so. In the words of this paper: *"The UKWIR report [on outcomes] noted that this consistency was ideal – in our view, it is essential".*

The priority attached to building upon, and refining, the significant changes introduced in PR14 should not, of course, preclude further regulatory innovation. But it might suggest a degree of caution to avoid change overload and a relatively high hurdle in relation to value added. Any changes also need to take account of the burdens they impose upon, and the incentives they create for, companies. Hence the need, identified earlier, to avoid generating work for companies that may, in the final event, not be required. So, if companies are asked to undertake research in the area of comparative performance commitments, the corollary should be that Ofwat is open to well evidenced arguments for some variation between companies.

More fundamentally, full account needs to be taken of incentive effects. The concept of dynamic incentives seems to me to require the most careful consideration. It is understandable, perhaps, that regulators might wish to see the benefits of innovation and efficiency passed onto customers as quickly as possible. But if this were, as looks likely, to undermine the incentives on companies to outperform, that could quickly work against the longer term interests of customers. It would also not be consistent with the real world operation of competition. In competitive sectors, companies strive to steal a march on their rivals by cost or product innovations that will give them advantage. That is how they make money for shareholders. The best companies retain those advantages for periods of time, as competitors struggle to match them. Certainly, there is no instantaneous market alignment to new cost or product standards of a sort that dynamic incentives would appear to be trying to replicate. Against this background, the existing five-year approach looks to be a tolerable approximation to real world circumstances, as well as having demonstrably provided the basis for past outperformance and the additional beneficial information this has provided to regulators at successive price reviews.

It will also be interesting to see how real world practicalities bear on whether, or how far, the regulatory use of comparators can extend outside the water sector. Ofwat is fortunate in having eighteen water companies from which to draw comparative data to assist in its regulatory judgements. Even that is not an easy task. Extension outside of England and Wales, or to other sectors, is likely to be more difficult. That was certainly my experience at the CAA. The issue for Ofwat will be whether the degree of judgment required to select and assess comparators makes interpreting the results of any such exercise problematic.

Several of the items identified above point to a broader issue for regulators: namely the extent to which the quest for more information, and more tightly defined controls and incentives, creates more complexity than regulatory value added. As Economic Insight's report identifies, there can be a difficulty in identifying whether apparent differences in data represent reality or 'noise'. To

the extent that the latter predominates that can make the regulator's job more difficult, even as the amount of information available increases.

This suggests the need for very careful evaluation of the case for change, from the viewpoint of both regulatory principle and implementable practice. This report represents a contribution to that assessment and offers, in a number of areas, pragmatic solutions designed to assist Ofwat design a process for PR19 that will help it build constructively on PR14.

Dr Harry Bush CB is a leading regulatory economist and policy advisor, who has held senior regulatory and economics positions over the course of his career.

Dr Bush was formerly the Group Director of Economic Regulation at the Civil Aviation Authority (CAA), where he was the Board Director with overall responsibility for the design, implementation and setting of price controls for London airports and air traffic control. Earlier in his career he held roles at HM Treasury (where he worked for 23 years) and specialised in issues relating to privatisation, regulation and supply side reform.

He is currently Vice-Chairman of University College London Hospital Foundation Trust (one of the UK's foremost specialist hospitals) and a non-executive director of NATS, the UK's regulated air traffic service provider. Dr Bush was educated at Oxford.

2. Executive summary

This report for Thames Water sets out our assessment of the key issues that must be addressed in order to determine an appropriate approach to outcomes at PR19. This executive summary provides a high level overview of the key themes of our report, our main recommendations and their implications.

In summary, our views are as follows.

(i) We think that having **clarity around the underlying principles and goals of the outcomes framework is the key to successfully building on the PR14 approach.**

(ii) Closely related to the above, a process, and evidence, that helps determine **where customer value genuinely varies across companies,** is critical.

(iii) Similarly, and building on the findings of the UKWIR report, we think **a consistent and concurrent approach to outcomes and cost assessment is essential.**

2.1. Building on the success of PR14 and the recent UKWIR report on outcomes

This report, prepared for Thames Water, provides an assessment of the key issues relating to the evolution of the outcomes framework for PR19. Building on the recent UKWIR study on outcomes,[1] our primary aim is to help support policy development in this area by: (i) providing evidence and analysis; and (ii) setting out the implications of this for the key choices facing Ofwat, companies and other stakeholders. The views and findings expressed here should also help inform engagement in Ofwat's forthcoming consultation on its approach to outcomes.

In undertaking this work, we recognise the undoubted successes of the outcomes framework at PR14. Across the industry there has been considerable support for its introduction. There is also an understanding that it contributed to there being a stronger focus on delivering against customers' needs and priorities. Accordingly, the objective for PR19 should not be to radically reform the approach per se. Instead, the focus should be on how to best retain the benefits it delivered, whilst also addressing some of the recognised issues that could make it even more effective.

Whilst our report addresses a range of issues currently under consideration by Ofwat and the industry relating to outcomes, it has two central themes, which we focus on in this executive summary.

Before setting out these themes, the following table provides definitions of the most relevant terminology (the main body of our report includes a fuller list of definitions).

> *"A process to determine where, and to what extent, customer value genuinely varies across companies is essential – as is a consistent approach to outcomes and cost assessment."*

Table 1: Selected definition of key terms

Performance commitments (PCs)	Refers to outcomes for which companies, within their Business Plans, 'commit' to a specified level of performance.
Bespoke PCs	Refers to outcomes, with an associated PC and Outcome Delivery Incentive (ODI) that are unique to the company proposing them in their Business Plan.
Common PCs	Refers to outcomes for which all companies are required to have an associated PC, with a consistent definition of the measure, and an ODI within their Business Plan (although the level of the PC itself could vary across companies).
(PCs subject to) Comparative assessment	Refers to *common PCs* whereby the level of the *PC* is determined through comparative (horizontal) analysis of performance across companies. Therefore, the level of the *PC* is the same across companies. These are a subset of common *PCs* (above).

Source: derived from a range of Ofwat documents. For full references see later.

2.2. Understanding when differences in the economically efficient performance level are 'real', or are 'measurement error'

Our first main theme starts from the fact that **outcome commitments are intrinsically tied to incentivising the economically efficient level of supply** (as recognised by the CMA in its redetermination of Bristol's price control).[2]

Once this point is understood, it is clear that - at the heart of the choice between bespoke and common PCs (and the application of comparative assessment) is the following

[1] *Setting performance commitments and incentives to deliver best value for money.' Frontier Economics, SYSTRA and Tynemarch – UKWIR report 16/RG/07/39 (2016).*

[2] *'A reference under section 12(3)(a) of the Water Industry Act 1991: A Report.' The CMA (2016); page 283.*

question: *for which outcomes is the economically efficient level likely to be the same or similar across companies, and for which is it likely to differ?*

Importantly, the efficient level of provision could vary across firms both due to variations in customer value, or the efficient cost of delivering outcomes. Accordingly, there is a need to develop and apply a clear framework for guiding the choice between bespoke, common and comparative approaches to setting PCs that reflects this, and other relevant considerations.

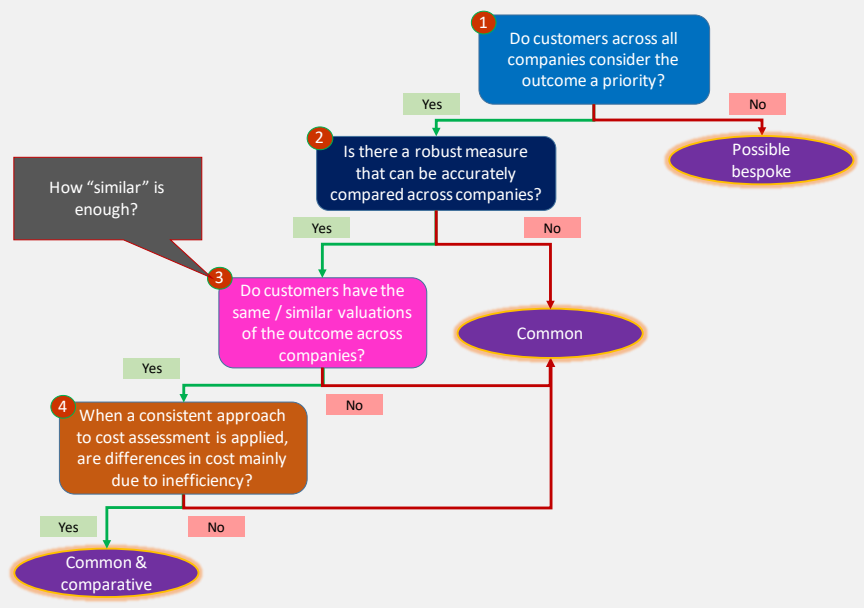
We have, therefore, proposed what we consider to be an appropriate framework, which is shown in the figure below.

At PR14, we would suggest that a key challenge was that there were not sufficiently well developed processes (and therefore evidence) in place to objectively determine where, and by how much, the economically efficient level of outcome performance varied, or was similar, across companies (i.e. stages 3 and 4 in our suggested framework).

When companies submitted their PR14 plans, Ofwat observed considerable variation in their outcomes proposals. Most notably, estimates of customers' valuation of outcomes differed substantially, meaning that companies proposed very different PCs for seemingly similar outcomes. In turn, this implied that customers of different companies could pay very different amounts for receiving similar performance levels.

To address the above, Ofwat undertook comparative assessment of a subset of outcomes proposed by companies. It then made 'after the event' interventions to company plans to set the level of PC for those outcomes based on its benchmarking (based on upper quartile performance).

Figure 1: Decision tree for choosing between common, comparative and bespoke approaches



Source: Economic Insight

The issue raised by the above is that *apparent* differences in either the marginal cost of provision, or customer value, across companies could reflect:

- **Genuine differences** in the economically efficient cost, or genuine differences in customer value. By this we mean factors that are outside of management control that drive cost differentials, which could include differences in the existing level of outcomes performance.
- **Measurement error** relating either to cost or value. Here we are referring a wide range of factors that could include: (i) in cost assessment, not taking into account factors outside of management control that might drive differences in efficient costs across firms; (ii) inconsistent recording of cost data; (iii) inconsistent definitions of outcomes measures across firms; and / or (iv) inconsistent recording of outcomes performance data – and so on.
- **Inefficiency** – namely, one company's costs are higher than another due to inefficiency; and / or a company is not delivering the efficient level of outcome performance.

At PR14, by imposing a consistent PC across companies for a subset of outcomes (using horizontal benchmarking as previously described) Ofwat effectively interpreted the observed variation *as being due to measurement error or inefficiency.* However, as explained, there was a limited basis on which one could objectively determine whether this was, in fact, the case.

Related to the above, one would never expect the economically efficient level of outcome performance to be *exactly* the same across companies. Consequently, the choice between common PCs and 'common PCs subject to comparative assessment', is a matter of degree. One might select a common PC where the economically efficient level of provision was expected to *vary materially* across companies, and apply comparative assessment where one expected it to be *similar.*

The above is central to our main findings in this area, which are that:

- The approach to outcomes at PR19 needs to **accommodate the inherent uncertainty** in determining the similarity of the efficient level of outcomes performance across companies.
- The process for developing the outcomes framework should nonetheless seek to **mitigate the uncertainty in order to achieve an incremental improvement over and above the PR14 approach.**

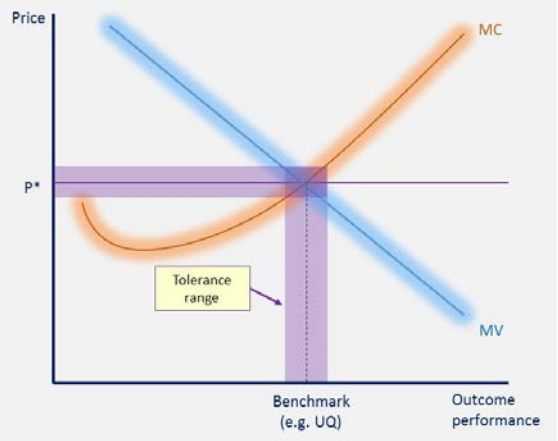
In addition to the above, an application of our proposed framework would logically imply that one would not expect there to be many common PCs, and even less would be comparative, at PR19.

2.2.1. An approach that can appropriately accommodate uncertainty

Relating to the first above bullet point, we advocate that there should be a tolerance range around any benchmark PC (for the subset of common PCs to which comparative assessment is applied). This will help mitigate the risk that Ofwat inappropriately adjusts company proposed PCs, resulting in consumer detriment.

This is illustrated in the following figure (see overleaf), where the principle is that the intrinsic uncertainty regarding the 'efficient' level of marginal cost (MC) and customers' marginal value (MV) would be strongly supportive of there being some flexibility, even where comparative assessment is applied.

Figure 2: Illustrating a tolerance range around common PCs subject to comparative assessment



Source: Economic Insight

In relation to the above proposal, we should highlight the fact that:

- It should be for companies to develop high quality evidence and analysis to justify any deviation from a benchmark PC – and they would be expected to engage customers and CCGs in this regard.
- It would be for Ofwat to determine the associated evidential hurdle (to be applied at Risk Based Review).
- That whilst the extent of 'tolerance' Ofwat allows would always be subjective, as a point of principle it would clearly be: (i) 'narrower' where there was greater certainty regarding the extent to which costs and customer value genuinely varied across companies; and (ii) wider where the opposite was the case. This is, therefore, strongly connected to our next main point.
- That, as we explain further below, because there is value in companies continuing to propose their own PCs even where comparative assessment is applied, this invariably implies that some 'tolerance' for variation from a benchmark should be permitted.

2.2.2. An approach that can mitigate uncertainty

Whilst the uncertainty described above is, to a degree, *inherent* we nonetheless consider that there is scope to achieve important incremental improvements for PR19.

In relation to the uncertainty regarding whether variations in customer value across companies are 'genuine', our view is that this can only be robustly addressed by applying a common set of customer research methods across the industry. This could both: (i) include innovative methods; and (ii) avoid undermining company's own individual customer engagement.

For PR19, however, the above may not be achievable for a variety of reasons. Consequently, we advocate a package to help mitigate uncertainty that would include:

- Ofwat identifying which common PCs will be subject to comparative assessment in advance of companies submitting their plans (just as Ofwat will identify common PCs).
- Ofwat also undertaking its comparative benchmarking analysis in advance of companies submitting their plans and, accordingly, making the details of its analysis (and the implied PCs) available – just as is the case in relation to cost assessment, where the benchmarking models are typically released to companies. This could then be used to support both customer engagement and CCG challenge.
- Companies could then propose their own PCs as part of their plans (including those where comparative assessment had been applied). As explained above, for 'comparative' PCs, company proposals *could* differ from the benchmark level, if Ofwat felt the evidence was sufficiently convincing. This is both for the reasons

explained above (i.e. to accommodate uncertainty) but also because, from a process perspective, were Ofwat not to allow any deviation, it would call into question why, in relation to PCs subject to comparative assessment, companies should develop their own research in the first place.

Uncertainty relating to the assessment of cost can be addressed by ensuring that there is a consistent approach to outcomes and cost assessment. This is the second main theme of our report.

2.3. The importance of a consistent approach with cost assessment

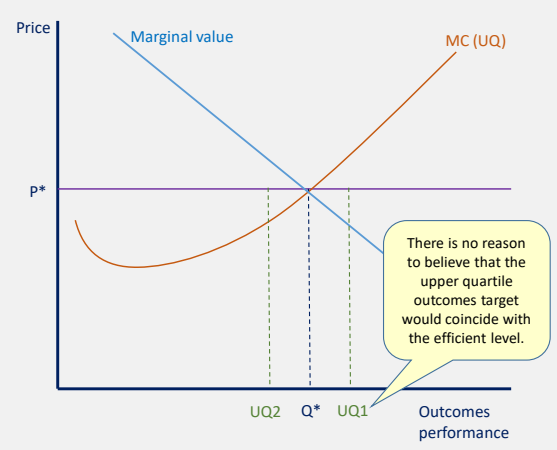
The second main theme of our report is the need to ensure there is consistency between the approach to (comparative) outcomes and cost assessment. By this we mean both intellectually and from a process perspective (i.e. the two should occur concurrently). The UKWIR report noted that this consistency was *ideal*[3] – in our view, it is essential.

The above again flows from the concept of economic efficiency. That is to say, because the 'benchmark' conceptually reflects a firm that is both 'technically' and 'allocatively' efficient, a benchmark selected for cost assessment must also represent a benchmark for outcomes performance – and vice versa. Put more simply: in competitive markets there are cost / quality trade-offs, and firms can adopt differing positions whilst still minimising their economic costs. For example, British Airways compared to EasyJet, say. This point recognised in economic theory, empirical studies, evidence and analysis and, importantly, by a range of economic regulators.

At PR14, because Ofwat's cost assessment set the cost benchmark at the upper quartile, in relation to outcomes, essentially *firms were funded for whatever level of outcome performance the upper quartile cost efficient firm could deliver.*

As noted by the CMA,[4] there is no intrinsic reason why this would coincide with the upper quartile outcomes performance for the industry. Consequently, and as illustrated in the following figure, the targets Ofwat set for those comparative PCs could have been below (*UQ2*) or, *more likely* above (*UQ1*) the performance level actually consistent with the funding provided.

Figure 3: illustration of the interaction between cost assessment and outcomes



Source: Economic Insight

The absence of a consistent and concurrent approach to outcomes and cost assessment risks customer detriment. Specifically:

- Companies may not have sufficient funding to deliver the outcomes customers genuinely want and are willing to pay for.
- Equally, companies might alternatively be 'over rewarded' for delivering outcomes.

[3] *Setting performance commitments and incentives to deliver best value for money.' Frontier Economics, SYSTRA and Tynemarch – UKWIR report 16/RG/07/39 (2016) page 18.*

[4] *'A reference under section 12(3)(a) of the Water Industry Act 1991: A Report.' The CMA (2016); page 283.*

2.3.1. Achieving consistency with cost assessment

The method by which one achieves a consistent approach is not straightforward – and we have considered the feasibility of various approaches.

Our initial work indicates that econometric methods (such as including outcome performance variables in the cost assessment models) may be problematic. At this stage, therefore, we suggest that an 'ex-post modelling adjustment' approach would be most practical. This could include either:

- adjusting the benchmarked level of PC (for any common PCs subject to comparative assessment) to reflect a level most consistent with upper quartile cost performance (if upper quartile remained the benchmark); or
- adjusting allowed costs to reflect the incremental cost of achieving outcomes performance consistent with the upper quartile firm.

Either approach should be underpinned by further analysis of the relationships between improvements in outcomes performance and incremental costs. This should include both top-down (statistical); and bottom-up, analysis.

This consistency of cost assessment also has practical implications that go to process elements of the approach to PR19. Specifically, they reinforce our observations about the benefit of making any benchmarking analysis relating to outcomes (that will be used to set incentives) available to companies in advance of plans being submitted – and that outcomes and costs should be considered concurrently.

2.4. Choice of benchmark

In addition to these two main themes, our report highlights a number of other important points relating to the evolution of the outcomes framework for PR19.

- There is a risk that the debate around 'static' and 'dynamic' approaches to benchmarking could be mischaracterised. In simple terms, it is perhaps easiest to think about this through the same lens that people debate the length of price controls. That is to say, a 'static' approach lets companies retain outperformance benefit (or suffer the consequence of under-performance) until the next price control is set. Consequently:
  - Static approaches deliver strong *technical efficiency* incentives (when one argues for longer price controls, this is typically a key motivation); whereas, on the other hand;
  - dynamic approaches allow the regulated price path to adjust more quickly to changes in actual performance (and therefore, might deliver better *allocative efficiency*).
- In the water industry, the fact that customer preferences and technology are not so fast changing, means one would expect customers to be best served by ensuring there are strong *technical efficiency* incentives for companies: ergo **a static approach seems appropriate.** Our analysis also shows significant year-to-year volatility in the outcome measures – which could be a concern, should dynamic approaches be applied.
- The choice of benchmark (i.e. frontier, upper quartile etc.) largely turns on 'how demanding' Ofwat wishes any challenge to be – balanced against ensuring the target is sufficiently achievable to provide a meaningful incentive. Our analysis points towards retaining an upper quartile approach (for the subset of

common PCs subject to comparative assessment). However, the choice between that and upper quintile appears finely balanced. This is an area where further evidence could be developed to help Ofwat reach a decision.

This report does not provide all of the answers in relation to outcomes at PR19. Nor, of course, will all stakeholders necessarily agree with our findings. However, the report does provide clarity around the key issues and identifies a number of important practical recommendations as to how the outcomes framework should be applied at PR19.

3. Introduction and aims

This chapter introduces our report, setting out our aims and objectives. In addition, we provide a re-cap of the PR14 outcomes framework, and highlight the key issues and challenges that emerged from this. We also briefly summarise the current direction of travel with regards to outcomes at PR19.

The key points highlighted in this introductory section are as follows.

(i) That the outcomes framework at PR14 was successful in a number of important regards. Most notably, it appeared to **genuinely drive a much greater focus on delivering outcomes aligned to customer priorities.**

(ii) The various issues and challenges that were encountered at PR14, in our view, can predominantly be traced back to **the need for clarity around the role and purpose of any outcomes framework** – and as such, this is the start point for our work.

(iii) Given the above, **the primary goal for PR19 should be to develop an approach that builds on PR14**, whilst also being practical within the time and resource constraints that exist.

3.1. The approach to outcomes at PR19 – key questions to address

This report for Thames Water sets out our assessment of the key issues that should help determine the approach to outcomes at PR19. The main aim of this paper, which builds on the recent UKWIR report on outcomes,[5] is to provide both practical recommendations and evidence to feed into Ofwat's forthcoming outcomes consultation; and ultimately, to help support the regulator in developing a methodology for PR19 in this critical area.

Our report is structured around addressing the following six questions:

- What is the appropriate **balance between 'bespoke' and 'common' performance commitments?** (and further, when should common performance commitments be assessed on a 'comparative' basis)?
- **How does the outcomes framework interact with cost assessment** – and how can consistency between the two be achieved?
- What is the appropriate **measure of the benchmark for comparative performance commitments?** (where here, by measure, we are referring to the choice between: upper quartile, upper quintile, average of best three, or frontier, etc).
- Should outcome **benchmarks for comparative performance commitments be set on an individual or aggregate basis?** (i.e. separate benchmarks for each comparative performance commitment, or one combined benchmark)?
- What are the relative pros and cons of **setting outcomes benchmarks on a static or dynamic basis?**

> *"The main aim of this paper is to help support the regulator in developing a methodology for outcomes at PR19."*

- What are the potential pros and cons of using **wider benchmarks** (e.g. international) as part of any outcomes framework?

In answering the above questions, we consider both: (i) the relevant *conceptual issues* that help inform what the 'right' approach should be; and (ii) the related *practical issues* that inform how the approach could be implemented in practice.

In order to arrive at answers to these questions that are both robust and internally consistent, it is necessary to consider, from first principles, exactly *what* the outcomes framework in the water industry is designed to achieve.

Consequently, before addressing these specific questions, in the next chapter of this report we firstly consider **the role and purpose of the outcomes framework.**

The subsequent sections of this report provide: (i) an assessment of the above questions; and (ii) our findings and recommendations. The remainder of this introductory section provides a brief re-cap of the PR14 approach to outcomes, and our views on the key issues arising from this; and (ii) an overview of the current direction of travel in relation to PR19.

The scope of our work excludes matters relating specifically to Ofwat's Service Incentive Mechanism (SIM).

[5] *'Setting performance commitments and incentives to deliver best value for money.' Frontier Economics, SYSTRA and Tynemarch – UKWIR report 16/RG/07/39 (2016).*

3.2. Definition of key terms

As there are many dimensions to the overall outcomes framework, it is helpful to start by defining a number of key terms, which we will make use of throughout our report.

These are summarised in the following table. The definitions provided are based on various Ofwat documents[6] and are, therefore, intended to align to the regulator's definitions.

Table 2: definition of key terms

Term	Definition
Outcomes	Refers to *what* companies propose to deliver for customers (and the environment) within their Business Plans.
Performance measures	Refers to the metric(s) used to measure performance in an *outcome* area (above).
Performance commitments (PCs)	Refers to *outcomes* for which companies, within their Business Plans, 'commit' to a specified level of performance. This is assessed via the relevant *performance measure* (above).
Comparative information	A subset of *performance measures* (with common definitions) that can be used - by Ofwat, Customer Challenge Groups (CCGs), customers and companies - to inform the comparative performance of companies across outcome areas. These need not have any associated *PCs.*
Bespoke PCs	Refers to *outcomes,* with an associated *PC* and *ODI,* that are *unique* to the company proposing them in their Business Plan.
Common PCs	Refers to *outcomes* for which *all companies* are required to have an associated *PC,* with a consistent definition of the measure, and an *ODI* within their Business Plan (although the level of the *PC* itself could vary across companies). These are a subset of *comparative information* (above).
(PCs subject to) Comparative assessment	Refers to *common PCs* whereby the level of the *PC* is determined through comparative (horizontal) analysis of performance across companies. Therefore, the level of the *PC* is the same across companies. These are a subset of common *PCs* (above).
Outcome Delivery Incentives (ODIs)	Refers to *outcomes* for which companies propose an 'incentive', based on their performance. The incentive can be reputational or financial. Where it is financial, the incentive can include penalties and / or rewards.
Willingness to Pay (WTP)	Refers to the 'value' that customers attach to receiving an outcome (typically defined as being the 'maximum' they will pay in exchange for the outcome).

** Bespoke and common PCs are sometimes referred to as bespoke and common outcomes*

[6] *'Setting price controls for 2015-20 – final methodology and expectations for companies' business plans.' Ofwat (July 2013); 'Final price control determination notice: policy chapter A2 – Outcomes.' Ofwat (December 2014); 'Reflections on the price review - learning from PR14.' Ofwat (July 2015); 'Water 2020: our regulatory approach for water and wastewater services in England and Wales.' Ofwat (May 2016); 'Outcomes – Water 2020 stakeholder workshop.' Ofwat (14 June 2016).*

3.3. Recap of the approach to outcomes at PR14

Appendix A provides a detailed description of the approach to outcomes at PR14, so here we briefly highlight a few key points:

- **Ofwat's new approach to outcomes at PR14 encouraged companies to focus on delivering against customer priorities.** Ofwat adopted a flexible approach that allowed companies to propose their own ODIs and related PCs.
- Ofwat's framework initially **included two common outcomes**[7] that would apply across all companies: SIM and leakage.
- In keeping with the above, **Ofwat's original intention was to play the role of 'reviewer'** in relation to outcomes, with company proposals assessed through its Risk Based Review (RBR) process.
- However, when company plans were submitted, **Ofwat had material concerns regarding their outcomes proposals** (mainly relating to PCs). Specifically, they implied customers would pay very different amounts for similar outcomes and performance levels; and that they did not appropriately balance risk and reward.
- **These concerns led Ofwat to de-scope ODIs from the RBR** – and prompted them to ask companies to resubmit more meaningful ODIs. This resulted in Ofwat ultimately following a very different process to the one originally envisaged (see figure below).
- **Ofwat made significant interventions in company proposed ODIs** at the draft and final determinations for PR14. These related to: setting ex-post comparative PCs across 5 outcome areas that were benchmarked on upper quartile performance[8]; calibrating company specific outcomes; implementing an aggregate cap and collar; and making adjustments to incentive deadbands.
- In relation to the above, we note that **Ofwat asserted that companies were funded for upper quartile service performance – but companies contested this point.**

Figure 4: Evolution of the PR14 process

Source: Economic Insight review of PR14 methodology, draft and final determinations

[7] *Which, at the time, Ofwat labelled as 'consistent outcomes' – see: 'Setting price controls for 2015-20 – final methodology and expectations for companies' business plans.' Ofwat (July 2013).*

[8] *Number of contacts from customers regarding quality of water; compliance with DWI water quality standards; number of sewerage pollution incidents; number of properties impacted by internal sewer flooding; and leakage.*

» For example, the following extract is taken from the submission of South East Water's CCG to the draft determinations:

"SEW initially advised the CCG that companies are not funded by Ofwat to achieve upper quartile performance because the totex modelling is based on average cost… the CCG can understand the Company's concern here. It is clear that the funding is for average performance (the models use historical average unit cost and apply it to activity forecasts for 2015-20, activity required to maintain levels of service) and the efficiency challenge applied is based on whether a company's costs represent upper quartile efficiency."[9] (i.e. lower than the average cost).

3.4. Key issues arising from the PR14 approach

There is widespread evidence that, in many important respects, **the overarching approach to outcomes delivered a number of benefits across the industry.** Ofwat's 'Lessons Learned' report (from summer 2015) found that:

"Overall, the sector viewed the introduction of outcomes as a positive development. There is almost a universal view from companies, customer representatives, other regulators and NGOs that the outcomes approach has:

- *created a much greater emphasis on the need for companies to understand and deliver the priorities of customers rather than those of the regulator; and*
- *given companies greater flexibility, alongside totex, to identify the most efficient way to deliver services that customers value most."*[10]

Following from the above, our view is that **it is important that any future approach to outcomes at PR19 retains the beneficial properties of the PR14 method.** Notwithstanding the above, there were a number of challenges associated with the application of the PR14 approach.

Ofwat's views on the key issues from PR14

Also contained in the Lessons Learned report, Ofwat set out some initial thoughts on areas where the outcomes framework could have been *more successful.* These included:

» That, echoing the concerns raised in the determination process, companies did not place 'enough' value at risk through ODIs, lessening their incentive power.

» That there was an inherent tension between encouraging companies to propose their own ODIs and PCs - and horizontal interventions (which were made retrospectively) in relation to the *comparative* assessment of outcomes. This, Ofwat has acknowledged: *"clearly runs counter to a key principle of PR14 and our direction of travel, which is to give companies the flexibility to identify and deliver the outcomes customers value most."*[11]

» Whilst Ofwat did not think that there was any contradiction between its approach to outcomes and long term planning (including asset health) the regulator suggested that perhaps more could be done to *"promote more structured and visible long term planning"*[12] through the outcomes framework.

---

9 *'South East Water Customer Challenge Group Submission to Ofwat on the Company's Draft Determination for 2015-2020.' (2nd October 2014).*

10 *'Reflections on the price review - learning from PR14.' Ofwat (July 2015); page 24.*

11 *'Reflections on the price review - learning from PR14.' Ofwat (July 2015); page 29.*

12 *'Reflections on the price review - learning from PR14.' Ofwat (July 2015); page 28.*

Key findings from the UKWIR report on outcomes

UKWIR recently published a report by Frontier Economics, SYSTRA and Tynemarch,[13] which examined options for developing the outcomes framework in relation to PCs and ODIs at PR19. Our report for Thames accordingly builds on this work. It is therefore worth highlighting some of the key conclusions it identified, as follows:

- **The PR14 approach to outcomes was not fundamentally flawed**, but incremental improvements require careful consideration.
- The methodology for PCs and ODIs needs to be developed in a timely way, recognising key interdependencies and trade-offs. Some of the specific issues highlighted in the UKWIR report included:
  - Consideration should be given to how costs and outcomes are assessed at the same time and specifically, whether cost allowances should be adjusted for outcomes.
  - Any methodology should be clear on whether, and how, different parts of the value chain contribute to specific outcomes (in the context of an increase in the number of price controls).
- There should be further research into the potential benefits of introducing longer-term commitments.
- The various options for measuring customer valuations should be explored – and any method should include clear guidance as to how companies should set incentive rates in the context of multiple estimates.

Our observations regarding issues raised by the PR14 approach

In addition to the considerations identified by Ofwat at the time, in our view the PR14 approach raises the following issues:

- The apparent tension between the freedom to use bespoke PCs, versus setting benchmarks based on comparative assessment, would seem to relate to the need for a clear framework to determine which method is most appropriate, why and where.
- Relating to the above, at PR14, even had such a *framework* existed, no *process* was put in place prior to determinations that could be used to make the above assessment. Nor, as a result, did Ofwat have access to the right kind of evidence to apply such assessments. For example, following company plan submission, clearly one could observe that WTP, proposed incentive rates, or *PCs* varied markedly across companies for relatively similar outcomes. However, it is not clear on what basis one could determine 'why'.
- In the absence of said process or evidence, Ofwat primarily addressed the above issue by making 'horizontal' interventions for comparative PCs, whereby it set upper quartile benchmarks. The implication of this would seem to be that Ofwat interpreted the variation in PCs across companies, for those outcomes, as being due to *measurement error or inefficiency* (where Ofwat's primary concern related to errors in measuring customer value, rather than cost).[14] In practice, however, this may, or may not, have been the case. For example, in unregulated sectors widely considered to be highly competitive, one observes firms with greatly differing

[13] *'Setting performance commitments and incentives to deliver best value for money.' Frontier Economics, SYSTRA and Tynemarch – UKWIR report 16/RG/07/39 (2016).*

[14] *Specifically, Ofwat's lessons learned report discusses concerns regarding variation in WTP estimates across companies in some detail. See: 'Reflections on the price review - learning from PR14.' Ofwat (July 2015); page 23.*

costs and service levels – yet those same firms could all be minimising their *economic costs* (for example, take Virgin Atlantic compared to EasyJet).

- As a direct result of the above, it is entirely feasible that some of the horizontal interventions may have been welfare enhancing (because the interventions *'rightly'* corrected for spurious differences across companies); whereas others will have given rise to customer detriment (because the interventions *'wrongly'* removed variation across companies that reflected genuine differences in customer value).
- There was an absence of any clear 'join up' to the approach to cost assessment. This is problematic because it might mean that, in some cases: (i) firms are *unable to provide the service levels* their customers genuinely want, because they are not sufficiently funded to do so; and / or (ii) may in some cases be delivering service levels *beyond* what their customers would want.

> *"There was an absence of any clear 'join up' to the approach to cost assessment [at PR14]. This is problematic because it might mean that firms are unable provide the service levels their customers genuinely want."*

- The lack of join up to cost assessment means it is not clear on what basis one could assume that firms *were* funded for delivering upper quartile outcome performance.[15]
- By requiring companies to achieve upper quartile performance across multiple individual outcome areas simultaneously, Ofwat actually set a more demanding outcomes target than had it set an 'aggregate' benchmark across multiple outcome areas (that is not to say that this was inappropriate, however).
- The fact that comparative assessments were implemented after plans were submitted may also have had a number of implications that merit consideration. For example, there may have been implications for:
  - *shareholder value,* because neither companies nor their investors therefore had an accurate view of plan value at the time of submission;
  - *regulatory risk,* because Ofwat ex-post changed its methodology;
  - *process,* because both companies and Ofwat became engaged in lengthy debates about the re-engineering of proposed ODIs; and
  - *future incentives for customer engagement,* because, in some cases, the ex-post horizontal interventions effectively 'replaced' company customer research.

It is important to recognise that the above observations are made with the benefit of hindsight. Indeed, one must keep in mind that the outcomes framework was new for the last price control and represented a step change for all stakeholders in the industry. Clearly, therefore, some of the issues that arose could not reasonably have been anticipated at the time – but rather,

[15] *Ofwat's PR14 cost models set a cost benchmark based on upper quartile that implicitly assumed 'average' service level performance (i.e. there was nothing in the modelling that explicitly controlled for 'upper quartile' service).*

emerged as the process itself evolved. It is also important to re-emphasise the benefits that the approach delivered, which are widely recognised.

Consequently, it is most pertinent to consider how, on a forward-looking basis, the benefits of the PR14 method can be retained, whilst adapting the approach to help address any identified issues. In our view, the 'common thread' that joins the various issues together is the need for as much clarity as possible around:

- what the outcomes framework is intended to achieve; and
- the related economics principles. This is something we therefore address in the next chapter of our report.

In the following we briefly summarise the current direction relating to outcomes at PR19.

3.5. Current direction in relation to PR19

Issues discussed in the December 2015 consultation document

In its December 2015 consultation document, Ofwat committed to retaining an outcomes focused approach at PR19, where a key aim was to *"build on the successes as well as learn the lessons from PR14."*[16]

The December document further set out a number of high level issues and principles that Ofwat would consider in relation to PR19. These included:

- **Encouraging a longer-term approach.** This included Ofwat considering whether there might be benefits in either encouraging or mandating certain outcome targets that span multiple price control periods.
- **Striking a balance between bespoke and common PCs.** Here Ofwat acknowledged that there would seem to be pros and cons associated with giving companies scope to innovate by proposing ODIs that reflect the requirements of local customers; versus having comparability that provides information with which to challenge company outcome commitments. Ofwat signalled that it would therefore specifically consider this matter through further consultation. Ofwat also raised the prospect that companies should submit the definitions (but not the related targets or incentives) for their PCs early in 2018 to help provide early clarity that could inform which outcomes might, or might not be, comparable.
- **Better incentives.** Here Ofwat raised the possibility of both allowing for outcome incentives to be paid 'within period'; and for revisiting the aggregate cap and collar, which the regulator said might *"dampen incentives for companies to improve their performance."*[17]

Policy positions on outcomes outlined in Ofwat's May 2016 decision document

In May 2016 Ofwat published its Water 2020 decision document. Within this, the regulator reaffirmed its commitment to retaining a strong focus on outcomes, based around a detailed understanding of what customers want. Ofwat also set out its 'preferred' policy position on a number of aspects of the outcomes framework for PR19. These included:

- In order to help encourage a longer-term focus, companies should provide long-term aspirations for all their PCs at PR19, and be transparent about these with their customers and other stakeholders.

[16] *'Water 2020: Regulatory framework for wholesale markets and the 2019 price review.' Ofwat (December 2015); page 142.*

[17] *'Water 2020: Regulatory framework for wholesale markets and the 2019 price review.' Ofwat (December 2015); page 160.*

- That companies should submit their definitions of PCs in advance of submitting their business plans. This is intended to allow Ofwat time to review the definitions and provide any feedback to companies ahead of them submitting their business plans.
- That a licence modification should be proposed that would facilitate (but not mandate) companies to have in-period ODIs.

In addition to setting out the above preferred policy positions, Ofwat's May 2016 document specifically asked for views on whether its approach to ODIs and PCs should be modified to embed a longer-term perspective. Here Ofwat set out four options:

- **Option 1:** the PR14 approach, where ODIs and PCs are set on a five yearly basis, reflecting the price control period;
- **Option 2:** as above, but where in addition companies are transparent with customers about their longer term targets for relevant PCs;
- **Option 3:** PCs are set on a longer term basis, with ODIs re-set at each price review;
- **Option 4:** implement long term PCs and ODIs (e.g. every 15 years).

The May document also further confirmed Ofwat's intended process for developing its approach to outcomes at PR19. This included:

Ofwat's recent workshop on outcomes indicated that the regulator was considering its approach to both how outcomes benchmarks were set – and applied.

- Confirmation of a separate consultation on outcomes in November 2016. Here Ofwat confirmed that the scope of this would include considering the balance between bespoke and common PCs; and the role of comparative assessment.
- An additional commitment to hold a more detailed consultation on the methodological aspects of the outcomes framework in July 2017. Ofwat indicated that the scope of this would include issues such as the methodology for calculating ODI rewards and penalties and the approach to setting dead bands, caps and collars.

Additional issues raised in Ofwat workshops on outcomes for PR19

In addition to its formal consultation and decision documents (summarised above) Ofwat has hosted industry workshops regarding the approach to outcomes at PR19.

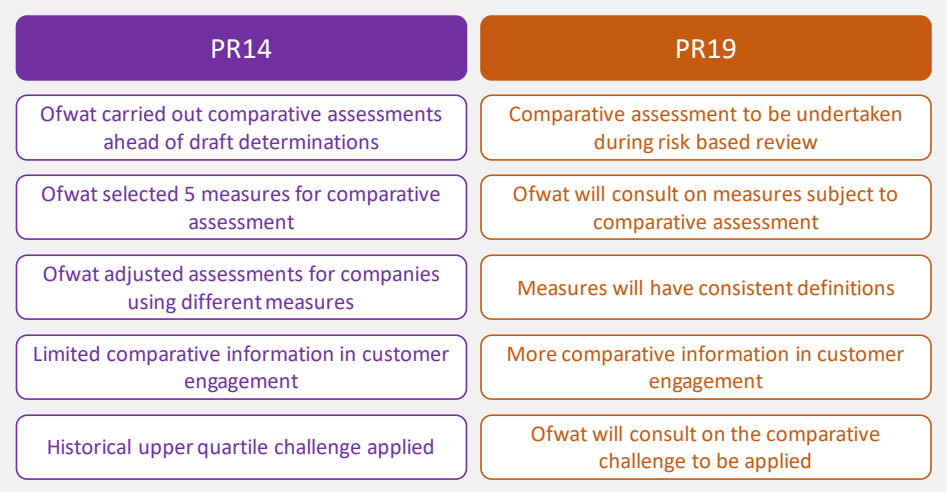
At a workshop hosted by Ofwat in June 2016,[18] further information was provided on the options the regulator is considering in relation to the *comparative assessment* of common PCs. This includes:

[18] *'Outcomes – Water 2020 stakeholder workshop.' Ofwat (14 June 2016).*

- Consideration of **the basis on which any benchmark should be set,** where Ofwat specifically tabled:
  - the continuation of an upper quartile approach;
  - an upper quintile approach;
  - average of the best three performers; or
  - frontier.
- Consideration of **'how' any benchmark is applied,** where options include:
  - static / historical;
  - forecast;
  - dynamic; or
  - dynamic (ratchet).
- Consideration of **the scope of how comparative PCs are applied**, where options include:
  - applying the benchmark to individual ODIs (as per PR14); or
  - applying the benchmarking to a basket of ODIs.

At the June workshop, Ofwat also set out the changes it envisages being made to the *process* for implementing any comparative assessment. These are summarised in the following figure.

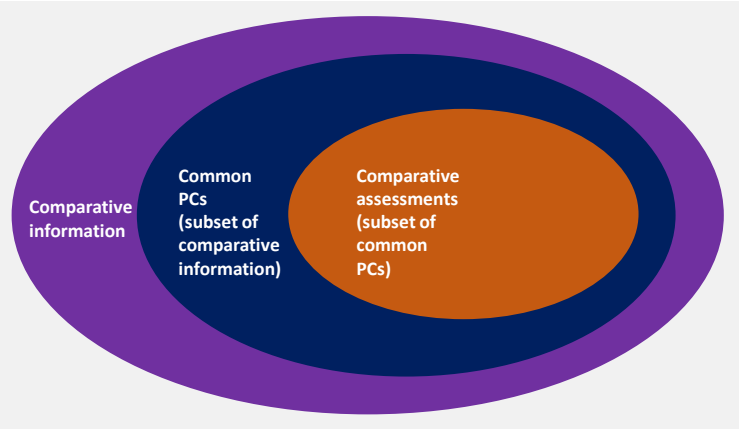
Figure 5 Proposed changes to Ofwat process for PR19



Source: Adapted from 'Outcomes – Water 2020 stakeholder workshop.' Ofwat (14 June 2016).

Relatedly, Ofwat provided further information regarding how it is thinking about the distinction between: (i) comparative information; (ii) common PCs; and (iii) comparative assessments (where the latter 2 are effectively subsets of the former – see the following figure).

Figure 6: Hierarchy of comparative information, PCs and comparative assessments



Source: Adapted from 'Outcomes – Water 2020 stakeholder workshop.' Ofwat (14 June 2016).

3.5.1. The PR19 timetable for outcomes

As the key aim of this report is to help move thinking forward to support the development of a framework and methodology for outcomes – it is important to be mindful of the practical considerations in this area.

This includes ensuring that any approach is consistent with the main milestones for PR19, summarised in Table 3 (right).

Table 3: PR19 milestones

Year	Date	PR19 Milestone
2016	May	Ofwat Decision Document.
	October / November	Ofwat consultations on outcomes, customer engagement and other design issues.
2017	June / July	Ofwat published methodology consultation for price review.
	December	Ofwat publishes final price review methodology.
2018	September	Companies submit draft business plans.
2019	March / April	Ofwat published determinations for enhanced companies.
	June / July	Ofwat publishes draft determinations for non-enhanced companies.
	December	Ofwat publishes final determinations.

Source: 'City briefing Water2020: Regulatory framework for wholesale markets and PR19.' Ofwat (December 2015)

Following from the above, where we set out the various options for how the outcomes framework might evolve to tackle the issues identified, we explicitly address the implications these have for key stakeholders. In particular, it is important to consider the sequencing of the activities required to deliver and apply a feasible outcomes framework for PR19. This involves the following:

- Companies will need to undertake customer engagement and research to determine both their prioritisation and valuation of outcomes. Whilst this is a continual and ongoing process for companies, the overall timetable and approach will need to allow for them being able to align the way they develop, evaluate (and then utilise) evidence to reflect Ofwat's final methodology.
- In relation to any approach to comparative outcomes, stakeholders will need to carefully consider when in the process it would be most desirable to provide clarity on precisely 'what' is being benchmarked and 'how' – and consequently, what data, information and analysis is needed (and by when) in order for this to occur.

4. The role of an outcomes framework

In this chapter we set out the role and purpose of an outcomes framework in the water sector. This starts from first principles, and then draws out the key implications for PR19 – which we subsequently explore in more detail in this report.

The key issues highlighted in this chapter are:

(i) That first principles provide a very **clear basis for identifying the aims of any outcomes framework.**

(ii) That the implied goals are **consistent with Ofwat's own stated aims, the CMA's interpretation of the outcomes framework, and Ofwat's duties.**

(iii) This, in turn, **provides a clear basis for addressing the key questions regarding how the outcomes framework should evolve at PR19** to build on the successes of PR14.

4.1. First principles of outcomes frameworks

In order to properly address the practical questions at the heart of refining the outcomes framework for PR19, it is first essential to provide clarity as to what the role and purpose of an outcomes framework in the water industry *should* be.

The appropriate start point for this is first principles – and specifically, the fundamental issues that economic regulation is intended to address.

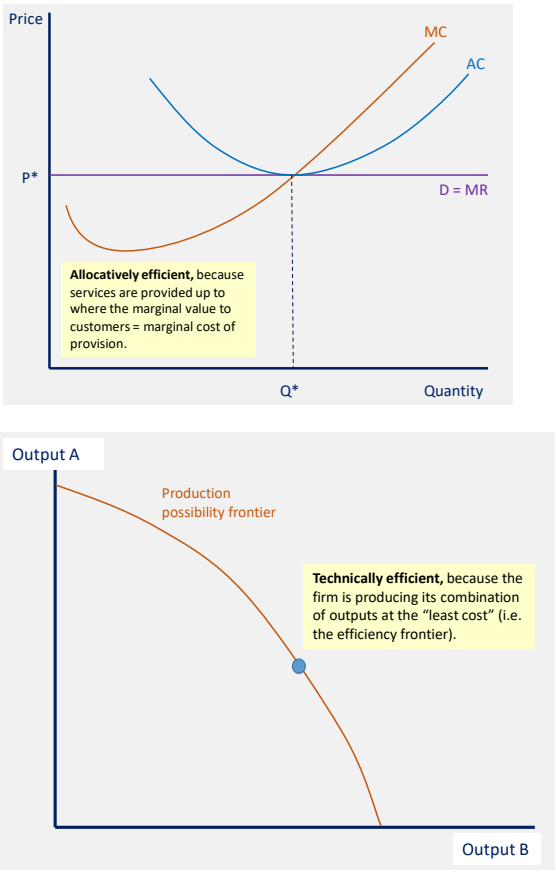
Economic regulation was originally intended to address market failures in naturally monopolistic industries

When modern day economic regulation first emerged in the 1980s, it was primarily in industries considered to be (in large part) naturally monopolistic. In simple terms, its aim was to correct known market failures associated with natural monopoly and, to some degree, to *simulate* competitive market outcomes. The rationale for this was that competitive markets deliver benefits to customers and society, because they are efficient. Specifically, in a competitive market firms:

- **Are Allocatively efficient.** That is to say, firms produce the *optimal* mix and quantity of outputs to reflect customers' preferences. Under an allocatively efficient outcome, firms will produce up to the point where the marginal cost of production is equal to the marginal benefit (or value) to customers. An allocatively efficient outcome is one whereby no one can be made any better off without making someone else worse off.
- **Are Technically efficient.** Here, we mean firms maximise the effectiveness of their means of production. In simple terms, they: (i) minimise their production costs for a given level of output; or (ii) maximise their output for a given level of production cost.

Allocative and technical efficiency are illustrated in the figure below.

Figure 7: illustration of allocative and technical efficiency



Source: Economic Insight

Absent intervention, we would expect natural monopoly to result in prices that are *above* the competitive level, and output (quality and / or quantity) that is *below* the competitive level. Therefore, economic regulation is designed to address these problems – and thereby incentivise firms to deliver prices and outputs that are more consistent with competitive markets.

Therefore, one might also think about outcomes frameworks as a means of 'simulating' competitive markets

Following from the above, first principles would suggest that, as part of a broader

regulatory approach, outcomes frameworks might be intended to help:

- Ensure that firms produce the outcomes customers want, at **the economically efficient level** (i.e. allocative efficiency).
- In conjunction with cost assessment, help ensure that the **outcomes are being delivered efficiently** (i.e. technical efficiency).

Over time, regulatory policy has evolved to focus more on 'promoting efficiency' in a broader sense

Over time, thinking with regards to economic regulation has evolved in a number of ways. This has included putting somewhat less weight on 'perfect competition' as a benchmark, because:

- Firstly, it will never be possible to accurately measure the efficient costs of production, or customer value. Hence, **attempts to 'second guess' what competitive market outcomes 'might be' could give rise to regulatory failures,** where the associated welfare loss might be greater than original market failure of the natural monopoly.
- Secondly, in any case **one does not observe perfectly competitive markets in the real world.**

For example, in a report for the CAA, Marshall and Robinson (2008) stated: "*A general danger in using perfect competition as a standard of comparison is that, since no market ever measures up to the standard, it can lead to inappropriate regulation… no regulator can know what the outcome of competition will be. A regulator who assumes it is possible to know what `efficient' prices are confuses the competitive outcome with the competitive process.*"[19]

Similarly, in commenting on the historical approach to energy regulation in Australia, Professor David Round (2002) referred to regulators adopting: "*a misguided application of perfect competition theory in the search for computational specificity and regulatory objectivity.*"[20]

Consequently, regulatory frameworks have evolved, so that the balance of emphasis has shifted further away from 'perfect competition' and more towards the **promotion** of the various forms of economic efficiency.

Importantly, this has increasingly incorporated approaches designed to help promote *dynamic efficiency,* which was not, historically, a main area of focus for regulators. Dynamic efficiency refers to the process whereby firms 'drive down' production costs over time, primarily through innovation and technology.

Examples of this evolution in regulatory approach include the following:

- Ofgem's development of the RIIO framework in energy was explicitly motivated by the recognition that the regulator's role should be to create incentives. Importantly, this included a strong focus on outcomes and on promoting dynamic efficiency.[21]
- In relation to telecoms, Ofcom's approach has explicitly recognised that is not appropriate or feasible to use perfect competition as a benchmark: *"anything close to the economist's concept of perfect*

[19] *'Comment on the Competition Commission Report: Stansted Quinquennial Review – Assessment of Competition at Stansted.' Marshall and Robinson (2008).*

[20] *'Assessment of Price Service Offerings.' As submitted to Queensland Competition Authority by Professor David Round (2002).*

[21] *For example, see 'RIIO Economics: Examining the economics underlying Ofgem's new regulatory framework.' Centre for Competition and Regulatory Policy; Dr Cloda Jenkins (2011).*

competition… was not likely to be achievable in telecoms."[22]

Importantly, Ofwat's reforms at PR14 (Future Price Limits) and its continued evolution in approach for PR19 (Water 2020) reflect many of the above observations. Ofwat's statement of principles for PR14 noted that: *"it is the economic regulator's role to mimic the effect of competitive pressure, and provide incentives for companies to operate in an efficient and customer-focused manner"*[23] (note the reference to competitive *pressure,* rather than perfectly competitive outcomes). Similarly, Ofwat's principles for Water 2020 explicitly include developing effective incentives and encouraging ownership and innovation to *"drive allocative, dynamic and productive [technical] efficiency."*[24]

Accordingly, first principles suggest that outcomes frameworks are also a tool for promoting efficiency

As regulatory frameworks have evolved more towards *'creating incentives to promote efficiency'* and away from *'simulating perfect competition'*, so it would seem sensible to think about the purpose of outcomes frameworks in the same way. Therefore, while the concepts of allocative and technical efficiency should remain highly relevant to the aims of an outcomes framework, one might think about their goals in terms of:

- creating incentives that *'make things better'* in a cost beneficial way; and further
- to ensure that '*dynamic efficiency'* is also considered in any approach.

There is discretion as to 'how much' weight one puts on the differing forms of economic efficiency in any outcomes framework

Once one understands that that the overarching goal of outcomes frameworks is the promotion of economic efficiency, first principles suggests that there will be *discretion* as to what weight one puts in the different forms of efficiency. This is because there are often trade-offs across these forms of efficiency, and so it is rarely possible to target all three without some form of compromise. Similarly, it is difficult to be definitive as to what the 'right' weight is for each form of efficiency – as this comes down to a judgement as to which is most likely to deliver the greatest benefits.

The objectives of an outcomes framework can also be thought of in terms of 'creating incentives' and / or 'monitoring or revealing information'

Finally, if one thinks about the 'goals' of outcomes frameworks in terms of economic efficiency, there is an additional dimension to their objectives that should also be considered. Specifically, one could develop various outcomes and measures for the purpose of:

- directly *incentivising* changes in behaviour that will enhance efficiency (where incentives could be financial or non-financial – i.e. reputational); and / or
- *monitoring and revealing information,* which itself could be of benefit to other aspects of the broader regulatory framework.

[22] *'Final statements on the Strategic Review of Telecommunications, and undertakings in lieu of a reference under the Enterprise Act 2002.' Ofcom (2005).*

[23] *'The role and design of incentives for regulating monopoly water and sewerage services in England and Wales – a discussion paper.' Ofwat (2010); page 9.*

[24] *'Towards Water 2020 – meeting the challenges for water and wastewater services in England and Wales: discussion document.' Ofwat (2015); page 42.*

Whilst, as noted previously, there may be trade-offs across the various dimensions of the 'aims' of outcomes frameworks, equally these aims are not mutually exclusive. Rather, they are matters of degree.

4.2. How Ofwat's stated aims and duties relate to first principles

The discussion set out above should be non-controversial. However, to demonstrate its relevance, it is helpful to consider how it relates to Ofwat's stated aims and approach – particularly in relation to outcomes more specifically. In the following we therefore briefly summarise some of the key evidence of relevance to this.

4.2.1. Ofwat's stated aims and objectives in relation to outcomes incentives

In developing its early thinking in relation to outcomes at PR14, Ofwat's stated aims were consistent with our first principles analysis. For example, Ofwat said: *"Focusing more on outcomes would also encourage efficiency. In particular, it should bring allocative and dynamic efficiency benefits."*[25]

Whilst ODIs were ultimately de-scoped from RBR at PR14, Ofwat's proposed RBR tests in relation to outcomes contained a number of features relevant to the above. For example, the RBR tests contained in the final methodology for PR14 included:

- In relation to the proposed level of performance (i.e. PCs) one test was that companies should demonstrate that performance commitments were in the interests of consumers, which Ofwat described in terms of being *"economic and efficient"*[26] (including over the longer term).
- As Ofwat explicitly intended for outcome commitment to be linked to customer value, the RBR tests also included criteria around the quality of WTP evidence, and the extent to which the proposed PCs and ODIs properly reflected customer value.
- A further RBR test relating to PCs was *'how reasonable are the costs'?* The combination of this, and the above test relating to customer value, is consistent with the notion that the framework was intended to promote economically efficient levels of service.

The above would seem to be highly consistent with the view that the PR14 outcomes framework was intended to promote efficiency, relative to some understanding of 'economically efficient' levels of outcome performance. Note, here we are not commenting on, in practice, the extent to which the final approach to outcomes as applied at PR14 adhered to first principles or not.

Ofwat has not yet set out in detail its aims or approach in relation to outcomes for PR19. This is for good reason, as a separate consultation process is in place in this regard. Nonetheless, Ofwat's May 2016 decision document did contain a number of statements of relevance to the issues highlighted here. These included:

- That ODIs and PCs will continue to be closely linked to the concept of 'customer value' (and, by implication, economic efficiency).
- Highlighting, as summarised earlier, the balance between bespoke and comparative PCs – where Ofwat explicitly referred to the potential for customer preferences and priorities to vary across companies. This is relevant to efficiency insomuch that it goes to whether 'efficient' outcome levels might vary across firms.

[25] *'Inputs, outputs and outcomes – what should price limits deliver? A discussion paper.' Ofwat (2011).*

[26] *'Setting price controls for 2015-20 – final methodology and expectations for companies' business plans.' Ofwat (July 2013) page 77.*

» A discussion of the potential role for longer-term commitments, which is relevant to the question of whether one is targeting efficient outcome performance in a more static or dynamic sense.

4.2.2. Evidence from the CMA's redetermination of Bristol Water's PR14 settlement

Consistent with our discussion here, in its redetermination of Bristol Water's PR14 settlement, the CMA took the view that the conceptual underpinning of outcomes targets (PCs) was that they reflect the economically efficient level of performance. *"The theoretical basis on which ODIs were designed appears to assume that the target is set at the economic level for the metric."*[27]

Furthermore, the CMA took the view that, in relation to comparative outcomes, there is no a priori reason to believe the 'upper quartile' (which Ofwat applied when benchmarking comparative PCs) will necessarily represent the economically efficient level.

4.2.3. How first principles relate to Ofwat's statutory duties

Ofwat's primary duties are laid out in the Water Industry Act (1991) and amended (2014), are as follows:

- *"To further the consumer objective.*
- *To secure that the functions of a water undertaker and of a sewerage undertaker are properly carried out as respects every area of England and Wales.*
- *To secure that companies holding appointments under Chapter 1 of Part 2 of the Act as relevant undertakers are able (in particular, by securing reasonable returns on their capital) to finance the proper carrying out of those functions.*

> *"The CMA's interpretation of outcomes performance commitments was that, conceptually, they refer to the economically efficient level."*

- *To secure that the activities authorised by the licence of a licensed water supplier and any statutory functions imposed on it in consequence of the licence are properly carried out.*
- *To further the resilience objective."*

As with any independent economic regulator, Ofwat has discretion as to how it carries out its functions in a way consistent with the above duties. Nonetheless, it is worth noting that:

» The historical interpretation of the 'consumer objective' (and relatedly, the promotion of competition) in the water sector is highly consistent with our prior discussion. Specifically, this aligns closely to the concepts of companies targeting outcomes that reflect consumer preferences at an efficient cost.
» The financeability duty is also very relevant, because historically this has been interpreted in terms of a notionally 'efficient' firm being able to earn its cost of capital. An outcomes framework that appropriately incentivised the various forms of economic efficiency – whilst also not undermining financeability, would, therefore, seem to align to this.
» The resilience duty is most relevant to: (i) potential trade-offs between static and dynamic efficiency when setting outcomes targets; (ii) the balance between current and future customers; and (ii) whether one is assessing economic 'value' and efficiency from a customer, or social (environmental), perspective.

[27] *'A reference under section 12(3)(a) of the Water Industry Act 1991: A Report.' The CMA (2016); page 283.*

4.3. The objectives of outcomes frameworks in other sectors

The broad 'aims' of outcomes frameworks, as implied by our first principles discussion, is also borne out by a review of the objectives of existing frameworks in place across a range of sectors. Earlier this year we provided a report to the ORR[28] examining how approaches to outcomes frameworks varied across sectors. We do not repeat the details of our findings here, but would highlight that we found clear evidence that frameworks could be classified in terms of:

- the extent to which they focused on promoting different forms of efficiency (we found they primarily focused on technical and / or allocative efficiency, although there were some examples of a more dynamic focus); and
- the extent to which they focused on **incentivising** behaviour or **monitoring.**

Generally speaking, where the regulator or body had some direct control over the setting of funding, and where customers or users paid for services, the regimes were more closely aligned to 'incentivising' (with an added consideration of the need to balance funding with outcome performance). Whereas, where the regulator / body did not directly control funding; and where customers did not directly pay for services, the frameworks tended to be more about monitoring.

4.4. Implications of economics first principles for the key issues that must be addressed for PR19

The matters described in this chapter have some clear implications for the key issues that need to be addressed in determining the approach to outcomes at PR19.

We expand on these in more detail within the next chapter of this report, where we provide our detailed assessment of each issue in turn. However, in summary the main implications for PR19 of a first principles consideration of outcomes frameworks are as follows:

- **The 'target level' of performance associated with an outcome (i.e. the PC) conceptually represents the 'economically efficient' level of performance.** However, in the same way that one should not necessarily be seeking to 'replicate' a perfectly competitive market outcome, the PC might be set to incentivise performance that is *closer* to the economically efficient level than would otherwise be the case without any outcomes-based incentives.
- In practice, **determining when the economically efficient outcome is 'the same' or is 'different' across companies is not straightforward** – particularly in relation to observed differences in customer value (as reported through WTP). As we noted in the introductory chapter to this report, at PR14 the absence of a process to help determine when observed differences were 'real' (as opposed to being due to 'measurement error') would seem to have been a key contributor to the extensive horizontal adjustments for comparative PCs.
- Following from the above, **where a 'comparative assessment' is applied**

[28] *'A cross sector review of outcomes frameworks: a report for the ORR.' Economic Insight (2016)*

that results in firms being set a benchmark level of performance for an outcome, one is implicitly assuming that the 'economically efficient' level of performance is the same across firms. This further implies that both: (i) customers' valuation of the outcome in question is identical; and (ii) that the economic cost of delivering that outcome is identical, across firms. Consequently, the balance between common and bespoke PCs in part turns on being able to develop evidence to determine where WTP varies and when it does not (as discussed above).

- **Where comparative assessment is used for outcomes, it is critical that the approach is consistent with cost assessment.** This is because, if appropriately identified, the 'benchmark' is, by definition, both technically and allocatively efficient.
- **The choice between different benchmarks is closely related to the challenge of accurately determining the frontier.** Just as it is not possible to 'second guess' what perfectly competitive outcomes might be, nor is it possible to observe the 'true' efficiency frontier for costs or outcomes. Therefore, whilst regulators wish to incentivise *improvements* in cost and outcome performance, it is recognised that the 'unadjusted' frontier is not the appropriate benchmark. Therefore, the choice between different benchmarks for common PCs, such as: frontier (with adjustment); upper quartile; upper quintile; etc, all reflect different ways of addressing this measurement problem.
- Questions regarding whether benchmarks should be set on a static or dynamic benchmark also relate to the need to exercise judgement regarding **'how much' weight to place on different forms of economic efficiency.**

Determining where the economically efficient outcome is 'the same' or is 'different' across companies is not straightforward

5. Assessment of key issues

In the following we set out our assessment of the key issues, structured around the questions that must be addressed in order to arrive at an appropriate outcomes framework for PR19.

For each question, we:

(i) Provide **a clear description of the key issues of relevance,** relating this back to our prior description of the role of the outcomes framework.

(ii) **Set out any relevant evidence or analysis** that can be used to help inform the appropriate choice of approach.

(iii) Identify the **practical implications for the PR19 approach to outcomes.**

(iv) Where appropriate, provide **our views as to what we think the most suitable options might be.**

5.1. The balance between bespoke and common performance commitments

In Ofwat's May 2020 Decision Document, the regulator highlighted that the balance between bespoke and common PCs will be a key consideration for its PR19 outcomes framework. Accordingly, this subject will be within the scope of Ofwat's forthcoming November consultation on outcomes. In the following, we set out our views and assessment of the relevant issues relating to this.

5.1.1. Defining bespoke and common PCs - and comparative assessment

Before discussing the factors that might influence the appropriateness of different 'types' of outcomes, it is first helpful to briefly re-cap what we mean by this (for full definitions see Table 2). Accordingly, here we are primarily discussing:

- **Bespoke PCs** are those that are unique to the *individual company* proposing them.
- **Common PCs** are those that apply to *all* companies (with a consistent definition of the performance measure).
- **Comparative assessment** refers to a sub-set of common PCs, where the level of the PC is determined by Ofwat, based on a benchmarking analysis of outcomes performance across companies (i.e. *horizontal analysis*).

At PR14, companies proposed a total of 171 outcomes (roughly an average of 10 per company). As shown in the following table, a sub-set of these outcomes were 'common' with associated PCs (7) and some were further subject to 'comparative assessment' (5). This classification is as identified by Ofwat.[29] This implies that, on average, companies proposed 3 bespoke PCs each.[30] We are aware, however, that the extent of outcomes, ODIs and PCs varied materially across companies (and that the average number of outcomes will differ for WaSCs and WOCs).

Table 4: Split of common and comparative PCs at PR14

Common	Comparative
• SIM; • Leakage; • supply interruptions; • water quality compliance; • water quality contacts; • pollution incidents; and • internal sewer flooding.	• supply interruptions; • water quality compliance; • water quality contacts; • pollution incidents; and • internal sewer flooding.

Source: Ofwat PR14 Final Determinations

5.1.2. Relevant economics and key issues

As we described in Chapter 4 of this report, any PC should reflect the economically efficient level. We also noted that the CMA explicitly interpreted the framework in this way in its redetermination of Bristol Water's PR14 price control.

Following from the above, the principles that underpin the approach to outcomes should reflect the following:

- Firms should deliver the **types** of outcomes customers want.
- That the **level / quality** of performance for each outcome should coincide with the *marginal value* customers attach to it and the *marginal cost* of providing it.
- That companies should be expected to **minimise their costs** of delivering the outcomes.
- That the above can also be considered in a **dynamic** sense, whereby firms drive shifts (improvements) in outcome performance over time.

[29] *'Final price control determination notice: policy chapter A2 – outcomes.' Ofwat (2014).*

[30] *Based on the fact that companies had 10 outcomes each, in total. Of these, every company must have had 7 common outcomes (a subset of 5 of which were subject to comparative assessment). Therefore, the residual (i.e. 10 less 7 = 3) must have been bespoke.*

The above provides a clear start point for considering the basis on which one would choose between bespoke, common and comparative PCs. In broad terms:

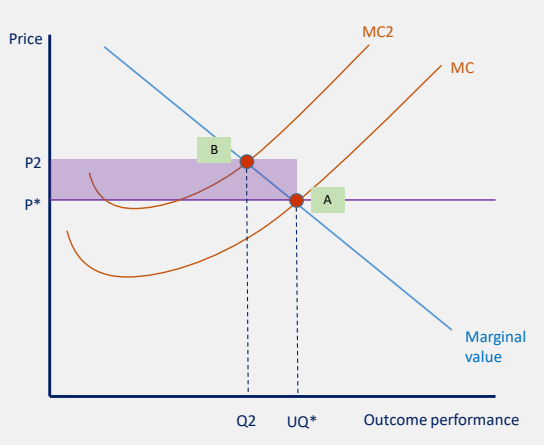
- The use of a **bespoke** PC would imply that the outcome is important to customers of the company in question, but not to water and wastewater customers more generally.[31]
- The use of a **common** PC would imply that the outcome is important to all water and wastewater customers,[32] but that the economically efficient level of provision *varies materially* across companies (which could also be consistent with different *measures* of the outcome being used across companies).
- The application of **comparative assessment** to common PCs would imply that: (i) the outcome is important to all customers; and (ii) that the economically efficient level of provision *is the same / similar* across companies (i.e. that customer valuation is the same and the economically efficient cost of provision is the same).

5.1.2.1. Differences in the economically efficient level of outcomes performance

Following from the above, the choice between a 'common PC' and a 'common PC subject to comparative assessment', turns on determining where the economically efficient level of outcome performance varies, or is the same, across companies. In the following we expand on why this matters – and the associated challenges.

Figure 8 illustrates the implications of differences in the *marginal cost* of providing an outcome across companies. Here, suppose there is a company with marginal cost of *MC*. The marginal value to its customers is shown by the blue line, which coincides with *MC* at point *A*, implying an appropriate outcome performance of *UQ**. (For simplicity we are assuming that upper quartile outcome performance is the efficient level. As we explain in section 5.2, this will not necessarily be the case).

Figure 8: Implications of difference in marginal cost of provision across companies



Source: Economic Insight

Now, suppose a second company has a *higher* marginal cost, shown by *MC2*. Here, the issue is that the 'gap' between the costs of the two companies could be due to:

- **Genuine differences in the economically efficient cost of provision.** By this we mean, factors that are outside of management control that drive cost differentials. Importantly, this could include differences in their existing level of outcomes performance.
- **Measurement error.** Here we mean factors that would genuinely drive differences in economic costs across the companies, but which cannot be observed or measured robustly (we discuss this further subsequently).

[31] *Or, indeed, that there is no evidence to the contrary – say because of incompleteness in customer research.*

[32] *In principle, if an outcome was a priority for customers of some, but not all, companies, Ofwat could create an additional category of PCs that sat between 'common' and 'bespoke' (i.e. they would be mandatory just for a subset of companies).*

» **Inefficiency.** Here we mean that firm 2 could simply be less efficient than firm 1, and so its costs are higher.

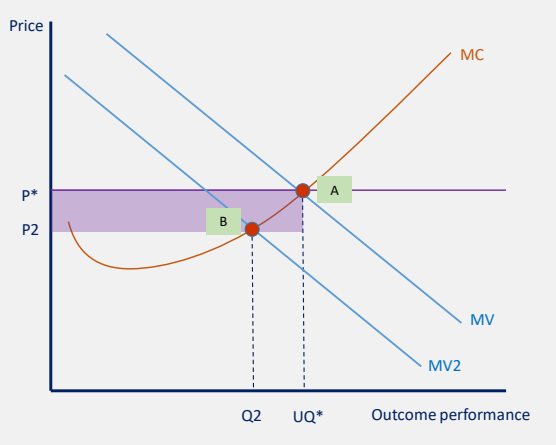
Importantly, the appropriate approach to setting PCs would vary, depending on which of the above factors drove the observed difference in marginal cost:

» If the difference was entirely due to inefficiency (or measurement error), then the economically efficient level of provision would be identical across the two companies, and so a single PC (*UQ**), would be appropriate (i.e. **one would apply comparative assessment** to a common PC).
» If the difference reflected genuine variation in the economic cost of provision, the appropriate response would be to allow company 2 a lower level of outcome performance, shown by *Q2,* consistent with point *B* in the figure (i.e. setting a single, benchmarked, PC using **comparative assessment would not be appropriate**).

As we explain in section 5.2, we think it imperative that Ofwat adopts a consistent approach to cost assessment and the comparative assessment of outcomes. By doing this, the industry will be better placed to determine precisely what is driving any observed differences in cost (and ensure consistency between funding and PCs). However, even were Ofwat to implement our recommendations in this area (as subsequently set out), some uncertainty would remain.

Similarly, the next figure illustrates the implications associated with potential differences in the *marginal value* customers attach to outcomes across companies.

Figure 9: Implications of differences in marginal value of outcomes across companies



Source: Economic Insight

Very simply, if Ofwat believed that the difference in marginal value across companies (i.e. the difference between *MV* and *MV2* above) was 'real', then it would be appropriate to allow the PC level to vary. If, on the other hand, Ofwat believed that company 2's estimate of value (*MV2*) was entirely due to inefficiency or measurement error, then setting a single PC (*UQ**) for both companies would be appropriate. Note in relation to *outcomes,* measurement error could include variations in the definitions of outcome measures, or indeed how measures are recorded.

Following from the above, the challenge faced by the industry at PR19 is how to address the uncertainty *inherent* in determining when variation in cost or value across companies is 'genuine', or is due to measurement error. Clearly, there is a risk of unintended consequences, which would harm customers, arising either from:

» Applying a single PC across all companies (comparative assessment) when, in fact, either or both the efficient marginal cost and / or customer value genuinely varies.
» Allowing PCs to vary across companies (a common PC without comparative assessment) when, in fact, the observed variation is due to inefficiency.

5.1.3. Implications for PR19

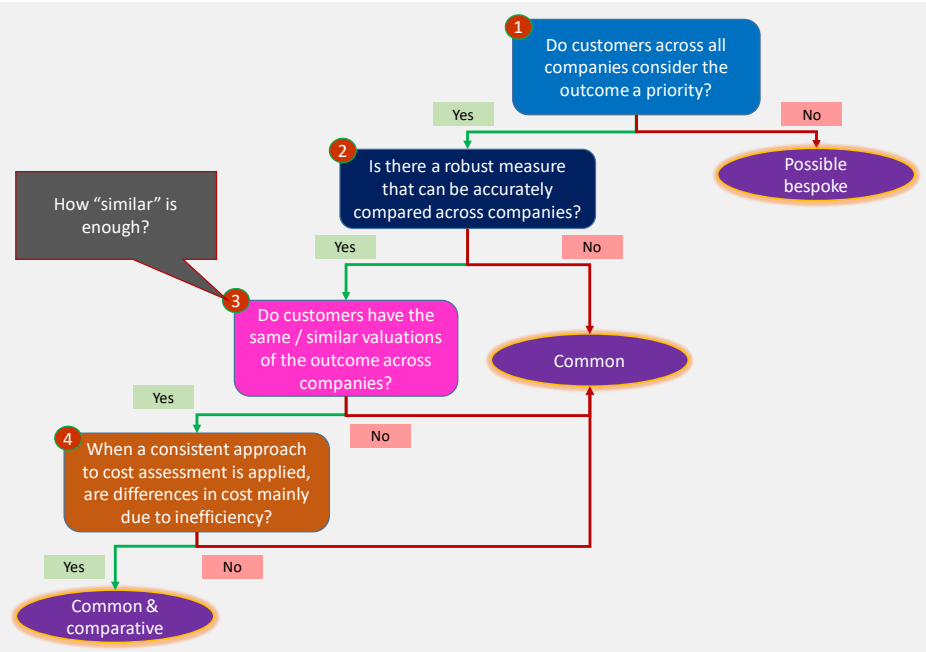
5.1.3.1. Bespoke, common, comparative decision tree

The first implication for PR19 arising from the previous discussion is that **a very clear analytical framework** is required to help underpin the choice between where and when bespoke, common and comparative PCs are appropriate.

We have therefore developed a decision tree, setting out our views as to the key questions that should determine these choices. This is shown in the figure below. The rationale for the stages of the decision tree are as follows:

- **Stage 1:** as a key principle is the concept of providing the outcomes that customers want, a *common* (rather than bespoke) approach, only makes sense where it is clear that the outcome is a priority for customers of *all* companies.
- **Stage 2:** whilst under *both* common and comparative approaches, one would wish to 'measure' the outcome in question, the need for a highly robust measure that can be compared like-for-like across firms is *essential* under a comparative approach (in order to ensure that the right incentives are being created).
- **Stage 3:** as the approach is intended to incentivise delivering performance levels in line with customers' marginal value of them, a *comparative* approach (which sets a consistent PC target across companies based on benchmarking) would only make sense where customer valuations are the same or, in practice, ***sufficiently similar*** across companies.
- **Stage 4:** the fact that the efficient level of outcome performance is also linked to the cost of providing it (marginal value = marginal cost) means that a comparative approach is appropriate only where (having applied a consistent approach to cost assessment, as explained in the next section of our report) one finds that any differences in cost are mainly due to inefficiency.

Figure 10: Decision tree for choosing between common, comparative and bespoke approaches



Source: Economic Insight

> *At the heart of successfully building on the PR14 approach to outcomes is the ability to identify and understand where and when customers' valuation of outcomes varies across companies.*

Ofwat's proposed principles for identifying 'common' PCs at PR19 are relatively consistent with our framework.[33] For example, Ofwat has said that common PCs are most appropriate where: (i) they are of particular importance to customers; and (ii) they are likely to be proposed by companies on the basis of customer engagement. Both of these could be read as being analogous to our Stage 1 question.

Similarly, Ofwat identified the existence of good quality comparable data as being a key principle for applying a common PC. As per Stage 2 of our framework, we think that the robustness and measurability of the outcome is perhaps most pertinent to the choice between 'common' and 'comparative' approaches.

A clear implication of the above is that, at the heart of successfully building on the PR14 approach to outcomes is the ability to identify and understand where and when customers' valuation of outcomes varies across companies. More broadly, success will critically turn on implementing a framework along the above lines – supported by robust evidence and data with which these questions can be answered with a high degree of confidence and credibility.

5.1.3.2. A flexible approach that can accommodate uncertainty to avoid unintended consequences

As we explained previously, the choice between common and comparative PCs turns on whether the economically efficient level of outcome provision is the same, or differs, across companies. In practice, however, it seems doubtful that one would *ever* expect the economically efficient level to be *exactly* the same. Accordingly, this would imply:

- **The choice between common and comparative approaches is, in fact, a matter of degree.** That is, a common PC is appropriate where one expects there to be ***material*** differences in the economically efficient level of outcome provision across companies.
- The use of **comparative assessment** to set a benchmark for a common PC is appropriate **where one expects the economically efficient level of provision to be *similar*** across companies.
- **Logically, one would not expect there to be many common PCs, and even less would be comparative** (which, by definition, are a subset of common PCs). This is consistent with Ofwat's provisional view that the number of common PCs should be "*relatively small,*"[34] in order to encourage company ownership of plans.
- Even where comparative assessment is applied, it may not be appropriate to set a single benchmark PC level for all companies. Instead, **an approach that allows for some (modest) variation is likely to be more suitable.** This is because, whilst we think there is scope to mitigate uncertainty (see below), it will remain an inherent feature of any outcomes framework. Consequently, imposing a single 'level' of PC across companies would risk unintended consequences that include deadweight loss arising from:
  - some companies receiving funding in order to provide a level of performance beyond that their customers actually

[33] *'Outcomes – Water 2020 stakeholder workshop.' Ofwat (June 2016).*

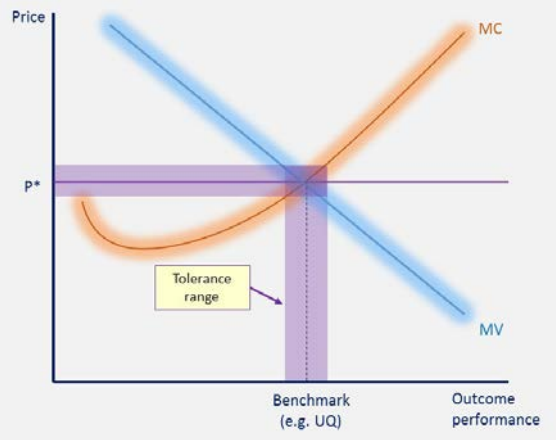
[34] *'Outcomes – Water 2020 stakeholder workshop.' Ofwat (June 2016).*

want (but will be forced to pay for); and / or
- other companies not being funded to provide the level of performance their customer demand (and, therefore, not actually providing that service level).

The above has the potential to create detriment for customers, companies and the environment.

Accordingly, **we would advocate there being a bounded 'tolerance range' around benchmarked comparative PCs,** as illustrated in Figure 11. The underlying principle of this is simply to reflect the fact that there will always be *some* uncertainty inherent in determining the similarity between marginal cost and marginal customer value across companies.

Figure 11: Illustrating a tolerance range around common PCs subject to comparative assessment



Source: Economic Insight

We discuss the process issues relating to the above subsequently. However, we would anticipate that key components of such an approach would include:

- Ofwat continues to undertake benchmarking for a subset of common PCs that it deems most suitable to comparative assessment.
- Ofwat retains responsibility for determining whether, and to what extent, it will 'allow' variation from its selected benchmark for PCs subject to comparative assessment.
- It would be for companies to make the case for any variation from the benchmark, and for Ofwat to determine an appropriate evidential hurdle.

The *extent* of any tolerance around a benchmark for comparative PCs is subjective. However, logically it is clear that the magnitude of this primarily turns on 'how robustly' Ofwat can distinguish between 'real' variation across companies; and variation that is due to measurement error. This, then, leads to the next key implication for PR19.

5.1.3.3. An approach that seeks to mitigate the uncertainty

We think the uncertainty described above is inherent – thus driving a need for a methodology that can appropriately accommodate it. Nonetheless, we also think the **process** for developing the outcomes framework at PR19 should explicitly seek to reduce the extent of uncertainty relative to PR14. There are two dimensions to this:

- **Ensuring consistency of cost assessment and outcomes,** so that there is a better understanding of the extent to which variation in costs across companies reflects inefficiency (and to ensure consistency of funding and outcomes). This is discussed in Section 5.2.
- Ensuring that the industry is better able to **distinguish between differences in customer value** that are 'real' as opposed to being due to measurement error. This is discussed below, within the next sub-section of our report.

> *"We would advocate there being a bounded 'tolerance range' around benchmarked comparative PCs."*

5.1.3.4. Implications for PR19 process

The above issues have some important implications for the *process* for developing and implementing the outcomes framework at PR19. Specifically, we think the issues go to **three** matters of process:

- whether Ofwat should **'pre-designate'** which PCs are comparative;
- **the means** by which Ofwat determines which PCs should be common / or subject to comparative assessment; and
- **the process** for comparative assessment itself (both in relation to outcomes and costs).

In the following we set out the various **options** relating to each of the above process matters. We then subsequently:

(i) set out how the options within each of the three process areas might be packaged together coherently (see Table 5); and

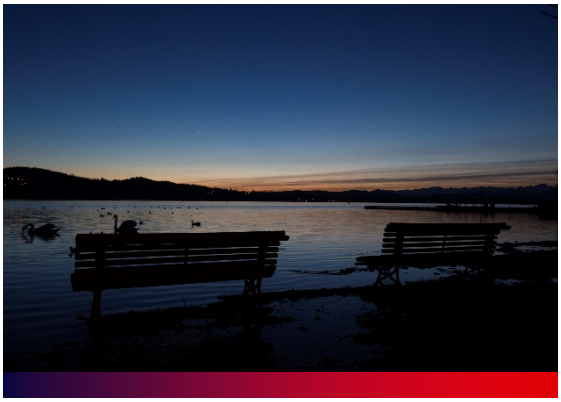
(ii) evaluate those **packages** to provide a recommended approach.

Process issue 1: whether Ofwat should 'pre-designate' which PCs are subject to comparative assessment

Firstly, under any approach, Ofwat will have to pre-designate which PCs are 'common'. This is because, by definition, a common PC is one that all companies are obliged to include in their plans. Clearly, companies can only do this if they know *which* outcomes this applies to in advance of submitting their plans.

There is, however, a choice as to whether Ofwat:

- also **pre-designates which common PCs will be subject to comparative assessment** before companies submit their plans **(Option A)**; or
- finalises this only **after companies have submitted their plans for PR19 (Option B).**



There would seem to be several advantages in Ofwat pre-designating which outcomes will be comparative in advance of plan submission

We note that Ofwat:

- is already committed to consulting on 'which' PCs should be subject to comparative assessment as part of its November consultation; and
- in its May 2016 decision document, Ofwat signalled its intention to ask companies to submit their PC definitions ahead of their Business Plans (Ofwat did not specify exactly *when* this should be, but noted that the earlier companies submitted definitions, the better).

In our view, there would seem to be several advantages in Ofwat pre-designating which PCs will be comparative in advance of plan submission. Specifically:

- it is likely to impact companies' approach to customer engagement – and so companies will be able to design and plan their engagement better, and more cost effectively, the earlier they have clarity on this;
- it is similarly likely to help ensure CCGs are as effective as possible, because they will have greater clarity as to the role they will play across the various proposed outcomes (for example, the nature of 'challenge' they would provide would clearly differ for an PC that was 'comparative' – and therefore subject to benchmarking by Ofwat –

relative to one that was common, supported by 'comparative information', but without any comparative benchmark); and
- by pre-designating outcomes, it is more likely result in an earlier definition of the measures of any comparative PCs – which Ofwat has acknowledged is desirable.

On the other hand, as a matter of process, if Ofwat wished to make use of company customer engagement data and evidence relating to PR19 to inform its selection of common and comparative outcomes in the first place, this might point to Ofwat not pre-designating which outcomes are to be comparative. This is closely related to the next point of process.

Process issue 2: the means by which Ofwat determines which PCs should be common / comparative

We earlier highlighted the fact that a key challenge with the PR14 approach was that companies' initial proposed ODIs implied (in some cases) very material differences in what customers should pay for similar levels of outcome performance. This, in turn, was a core reason for Ofwat's horizontal interventions to benchmark performance across 'comparative' PCs (based on upper quartile).

Following from the above, we would suggest that Ofwat's horizontal interventions would imply that the regulator interpreted the *variation* across companies in these areas as being due to 'measurement error' or inefficiency, rather than reflecting genuine differences in the economically efficient level of provision. In practice, however, there was limited information with which one could objectively say with any certainty in what circumstances this was, or was not, the case.

In hindsight, what this illustrates is the critical importance of having:
- high quality data and evidence to inform such decisions; and therefore,
- a process both for collecting the required data and evidence and for making the decisions.

From a process perspective, we consider that there are four broad options for addressing the above – as follows:

» **Option A: Ofwat (or other stakeholders) commission cross-company customer research.** This is to ensure that a consistent set of valuation methodologies are applied to all companies for PCs that could potentially be 'comparative'. The rationale for this process is that it could be designed in such a way that the resultant WTP estimates could control for 'company' to see if there are any statistically significant differences in customer valuations. That is to say, once the research is complete, regression analysis could be undertaken (which could, for example, include dummy variables for 'companies') to test whether, and to what extent, customer value varies across them. Importantly:
- **This approach does not rule out using more innovative methods**, such as experimental or behavioural techniques to inform WTP, alongside stated preference. Rather, the principle is simply that a consistent method, or set of methods, is used to ensure that variations in reported WTP are themselves *not because of differences in method.*
- **This approach does not undermine the importance of companies taking forward their own customer engagement work.** For example, under all options, company specific engagement and research would remain essential in relation to any bespoke PCs. In addition, it might still be beneficial for companies to undertake their own engagement and

research into comparative PCs – but only if Ofwat might allow some *deviation* from any comparative benchmark it set (say through the RBR process). We explore this second point subsequently.

- **Option B: Ofwat could make use of the existing WTP evidence base from PR14, and undertake a more detailed 'backwards looking' evaluation of the results.** This approach would involve a 'deep dive' into the differences in methodologies, definitions and so on, so that Ofwat could better understand, across the various outcome areas, the extent to which any reported differences in customer value were 'genuine' as opposed to being driven by other factors. This, in turn, would put Ofwat in a stronger position to identify which PCs should be common / comparative (note, this approach does not imply one would use the PR14 WTP data to set PCs, incentive rates and so on at PR19 (which we discuss under 'process issue 3 subsequently). Rather, here we are focused only on 'how' one should determine which outcomes should be common / comparative and so on in the first place).
- **Option C: as above, but using the WTP evidence provided by companies to support their PR19 plans.** In our view, this would be likely to be more effective if Ofwat: (i) gave guidance in advance, specifically aimed to encourage greater consistency in how customer value was assessed – which would seem to also include clarity around definitions; and (ii) Ofwat also set out clear criteria that would be applied to evidence to help determine its robustness and comparability. Note, this is closely linked to the process for comparative assessment itself, which we discuss subsequently – and to the previous discussion of whether Ofwat pre-designates comparative outcomes (i.e. by definition, Ofwat could not pre-designate under this approach).
- **Option D: more qualitative based approaches.** This would likely involve drawing on a range of evidence to help inform a 'judgement' as to for which outcomes one would intuitively 'expect' customer valuations to be similar across companies.

There are pros and cons of the various processes outlined above. At a fundamental level, however, it is difficult to see how one can robustly resolve the question of whether variations in customer valuation are 'genuine', or due to other factors, without a consistent customer research methodology being applied across the customers of all companies. We have also explained that, if applied appropriately, this should not undermine the importance of company-led customer engagement. Over the longer-term, therefore, there would seem to be strong merit in considering developing a process to deliver this.

From a PR19 timetable perspective, however, the feasibility of implementing this process would seem to be challenging.

Were one to rule out the first option due to feasibility, this would point to:

- using a combination of PR14 data and qualitative analysis to inform the choice of which PCs should be 'comparative' if Ofwat wished to pre-designate this; or
- using WTP data and evidence from company PR19 plans, were Ofwat willing to delay finalising which PCs would be 'comparative' *until after plans*

were submitted.[35] However, this approach would seem to reduce the ability of the CCGs to effectively challenge Business Plans.

We note that the recent UKWIR report characterised the above discussion in terms of 'customer engagement', where the paper identified two options: (i) 'improved WTP', which involved continuing with company specific WTP research as per PR14, but with tools to improve its quality and consistency – UKWIR noted that this could include trials and experiments; or (ii) a 'single generic survey' applied across companies. The UKWIR report is, therefore, somewhat consistent with the process options we have identified – where the clear motivation is to help better unpick reported differences in customer value across companies.

However, in our view, the options presented in UKWIR perhaps imply a 'starker' choice than the one facing the industry in reality. Specifically:

- it is not clear why a consistent research method for evaluating WTP across companies in relation to certain outcomes would 'rule out' more innovative methods (i.e. they just need to be 'the same'); and
- a consistent research method could sit alongside company-led research (i.e. these are not an either / or).

Importantly, which of the four options is implemented for PR19 would have **implications for any 'tolerance' range for common PCs subject to comparative assessment.** Specifically:

- Option A (a consistent cross-industry research method) would seem to have most scope for reducing uncertainty as to explaining observed variation across companies. Consequently, adopting this approach might allow Ofwat to set a narrower tolerance range for comparative PCs – or potentially, not have a tolerance range at all (depending on its confidence regarding cost assessment).
- On the other hand, the alternative options all carry a higher degree of uncertainty – and so would reinforce the need for a tolerance range; and for that range to be 'wider' than under option A.

***Process issue 3:* undertaking comparative assessment**

Having: (i) decided whether to notify companies in advance as to which PCs will be subject to comparative assessment (process issue 1); and (ii) determined how it will identify which PCs comparative assessment will be applied to (process issue 2), Ofwat will further need to decide 'how' the comparative assessment itself will be undertaken.

As noted previously, at PR14 Ofwat applied 'horizontal adjustments' to certain PCs ex-post (after companies had submitted their plans – and, in fact, outside of the RBR process). As we explained, these forms of intervention were not consistent with the way in which Ofwat *originally* envisaged the process being applied (i.e. Ofwat originally intended to be an 'overseer'). However, it is important to stress that, having reviewed the original PCs proposed by companies, it is understandable that Ofwat had concerns; and why, therefore, the regulator felt it necessary to make these interventions.

Whilst there was a clear rationale for Ofwat to intervene, the fact that this was not signalled in advance would seem to give rise to some issues that should be addressed going forward. For example, the

[35] *Unless the process were modified such that companies provided their PR19 related WTP evidence in advance of plan submission).*

retrospective nature of the changes could have negative connotations in terms of: regulatory risk, plan value and, therefore, customer bills over the longer term.

Consistent with the above, Ofwat has acknowledged[36] that it would be beneficial if, for PR19, a clearer process was developed in good time – and made transparent to the industry. We therefore agree strongly with Ofwat on this point.

In the following we therefore consider three options for the process by which Ofwat could *undertake* comparative assessment.[37]

- **Option A:** The comparative analysis and the PCs it implies (as determined by whatever benchmarking method it deems appropriate), are **published by Ofwat in advance of companies submitting their plans. Companies do not propose their own PCs** for these comparative PCs, as Ofwat does not intend to allow any deviation from its benchmarks.
- **Option B: As per Option A, but where companies do also propose their own PCs** for the comparative PCs. Accordingly, Ofwat may allow some deviation around its benchmarks – and this would be determined at RBR, based on the quality of company evidence. As this option allows for the possibility for deviation around the benchmarks, Ofwat *could* also publish its own proposed 'bounded range' as an output of its comparative assessment.
- **Option C:** that the comparative analysis and implied PCs are **published by Ofwat during the RBR process – after companies have submitted their plans**. For reasons we explain below, under this option some 'deviation' from Ofwat's benchmarked PCs would be allowed and would be determined at RBR, again based on the quality of evidence provided by companies.

The key differences between the approaches are that:

- Under the first two options A and B, companies would have earlier visibility of Ofwat's comparative analysis, and the PCs this would imply. The only distinction is whether companies propose their own PCs *at all* in the first place.
- Under Option C, companies would continue to propose their own PCs for comparative outcomes. However, as they would have no prior sight of Ofwat's benchmark or comparative analysis, one might anticipate there being a greater level of variation in company's initial proposals (relative to Option B). This would, therefore, seem to make more sense if:
  - Ofwat intended to be more 'permissive' of variation in PCs in the first place (as otherwise, interventions at RBR to limit or remove variation across companies would simply seem to be a more convoluted way of achieving the same objectives as the other options); furthermore
  - Option C might also seem superficially attractive if Ofwat intended to use 'forecast' information to set the benchmark – and where it felt this information might be of benefit. In practice, however, the value of this would be limited, as it would seem to create incentives for companies to 'game' their forecasts (a problem that cannot arise in relation to historical data). In addition, forecast information would be inherently less accurate than 'actual' performance data.

[36] *'Outcomes – Water 2020 stakeholder workshop.' Ofwat (June 2016).*

[37] *As previously defined, comparative assessment will apply only to a subset of common PCs.*

> *"The choice of process for the comparative assessment of outcomes is linked to cost assessment. This is because, as we explain subsequently, there is strong merit in considering the two concurrently."*

We note that the choice of process is also linked to cost assessment. This is because, as we describe subsequently (and in line with the UKWIR report) there is clear merit in addressing any benchmarking of outcomes and cost assessment concurrently, and in a consistent manner.

Relatedly, regulatory precedent is typically that companies have prior sight of cost benchmarking models and analysis. Indeed, for PR19 Ofwat has already committed to providing details of: *"our cost assessment approach and cost-efficiency modelling, as part of the price review methodology statement which could inform companies' development of their view of efficient costs to deliver business plans and items such as special factor claims."*[38]

The philosophy underpinning this would seem to suggest adopting a similar approach to the process for outcomes benchmarking.

5.1.4. Evaluating packages of options

Ultimately the three process dimensions described above are all highly inter-dependent. Consequently, when assessing their pros and cons, it is important to consider them simultaneously. Given this, in the table overleaf and subsequent discussion, we summarise and evaluate a set of three cogent packages.

[38] *'Water 2020: our regulatory approach for water and wastewater services in England and Wales.' Ofwat (2016); page 38.*

Table 5: Packages of process options

Key process related questions	Process issue 1 Whether Ofwat pre-designates which PCs are subject to comparative assessment	Process issue 2 Basis for determining 'similarity' of customer valuations across PCs that could potentially be subject to comparative assessment	Process issue 3 Undertaking comparative assessment
Option A	Ofwat pre-designates outcome classifications in advance in plan submissions.	Cross-company consistent methodology applied.	Ofwat publishes comparative assessment and implied PCs in advance of plan submission – companies do not propose their own PCs (i.e. PC = benchmark).
Option B	Ofwat designates outcome classifications after plan submissions.	Detailed review of PR14 data and evidence to unpick differences in WTP.	As above, but companies do propose their own PCs (i.e. PCs could deviate from the benchmark, to be determined at RBR).
Option C		As above, but using PR19 data (with guidance and criteria given in advance to improve quality and comparability).	Ofwat provides no sight of comparative assessment or PCs – prior to plan submission. (Implies Ofwat is more permissive of deviations from the benchmark at RBR and / or intends to use forecast information).
Option D		Qualitative approaches.	
Packages of process options for consideration			
Package 1	Option A	Option A	Option A
	• Ofwat pre-designates PCs subject to comparative assessment in advance of plan submissions. • A cross company methodology to determine which PCs are most suitable to comparative analysis. • Ofwat publishes comparative assessment and implied PCs in advance of plan submission – no company proposed PCs for those comparative outcomes.		
Package 2	Option A	Option B	Option B
	• Ofwat pre-designates PCs subject to comparative assessment in advance of plan submissions. • Ofwat uses PR14 data and qualitative information to determine which PCs are most suitable to comparative analysis. • Ofwat publishes comparative assessment and implied PCs in advance of plan submission – companies continue to propose their own PCs for comparative outcomes – some deviation from the benchmark allowed at RBR, subject to high quality evidence.		
Package 3	Option B	Option C	Option C
	• Ofwat designates outcome classifications after plan submissions. • Ofwat uses WTP data from PR19 plans to determine which PCs are subject to comparative assessment. • Ofwat publishes comparative assessment and implied PCs after plan submission. Ofwat is more permissive of deviations from its benchmark (relative to Package 2).		

Source: Economic Insight

Initial evaluation of the packages

We think that the advantages of Ofwat pre-designating which PCs will be 'comparative' are compelling – and indeed, would seem to have several process related advantages.

We also think that, ultimately, the question of whether – and to what extent – customers' valuation of outcomes varies materially across companies can only be addressed by applying a common method(s) across companies (as per Package 1). This, then, should be the longer-term aspiration. For the purpose of delivering PR19, however, we are mindful that it might be more practical to determine between comparative and common PCs using the existing evidence base (as per Package 2).

Finally, we are conscious that, even with improved evidence, Ofwat's ability to precisely identify the appropriate benchmark will be limited – and so imposing a single benchmark with no scope for deviation (as per Package 1) may be undesirable. It would also undermine company Plan ownership as, logically, companies would not propose their own PCs for comparative outcomes. In relation to the choice between the remaining packages, we think:

- Package 2 offers some advantages, because – as companies would continue to propose their own PCs – the RBR would provide a good forum through which Ofwat could consider whether company evidence suggests there were some valid reasons why the (comparative) PC should deviate from the benchmark, and could therefore be permissive of this. For example, current outcome performance might, in part, reflect differences in the historical pattern of investments across companies – or might reflect certain regional factors. The point, however, is that this approach provides Ofwat with *discretion* to consider this (subject to it setting a hurdle for high quality evidence) and, unlike Package 1, retains a rationale for companies to propose their own PCs in the first place. Importantly, therefore, this should reinforce the principle of 'company owned plans'.
- Package 3 would seem relatively unattractive because it only makes sense if Ofwat was willing to be permissive of *greater* variations in the PCs than under Package 2. In addition, it would seem to reduce certainty around the development of company Business Plan development.

Accordingly, **the analysis set out here would point to an approach based on 'Package 2'.**

Role of the CCGs

We would emphasise that, under all of the packages we have considered, CCGs would continue to play a critical role, consistent with that envisaged by Ofwat. Specifically, we would fully anticipate that, as indicated in Ofwat's May 2016 decision document, the role of CCGs would include – under all eventualities:

- **Providing independent challenge** to companies.
- **Providing assurance** to Ofwat regarding: (i) the quality of a company's customer engagement; and (ii) the extent to which the results of this engagement are driving decision making and are reflected in the company's plan.

The precise process selected would, however, impact the balance of where one would expect CCGs to focus their efforts. For example, if Ofwat selected a process whereby it 'set' the PCs for comparative outcomes in advance based on benchmarking - with no company proposed PCs as per package 1- clearly CCGs' challenge role would not be relevant to

those specific comparative outcomes. Alternatively, if Ofwat published its comparative analysis and implied PCs in advance, but allowed companies to propose their own with some scope for variation (as per Package 2), clearly the challenge role for those outcomes would remain – and so on. Note, the 'challenge' role of CCGs would remain for common and bespoke outcomes in all cases.

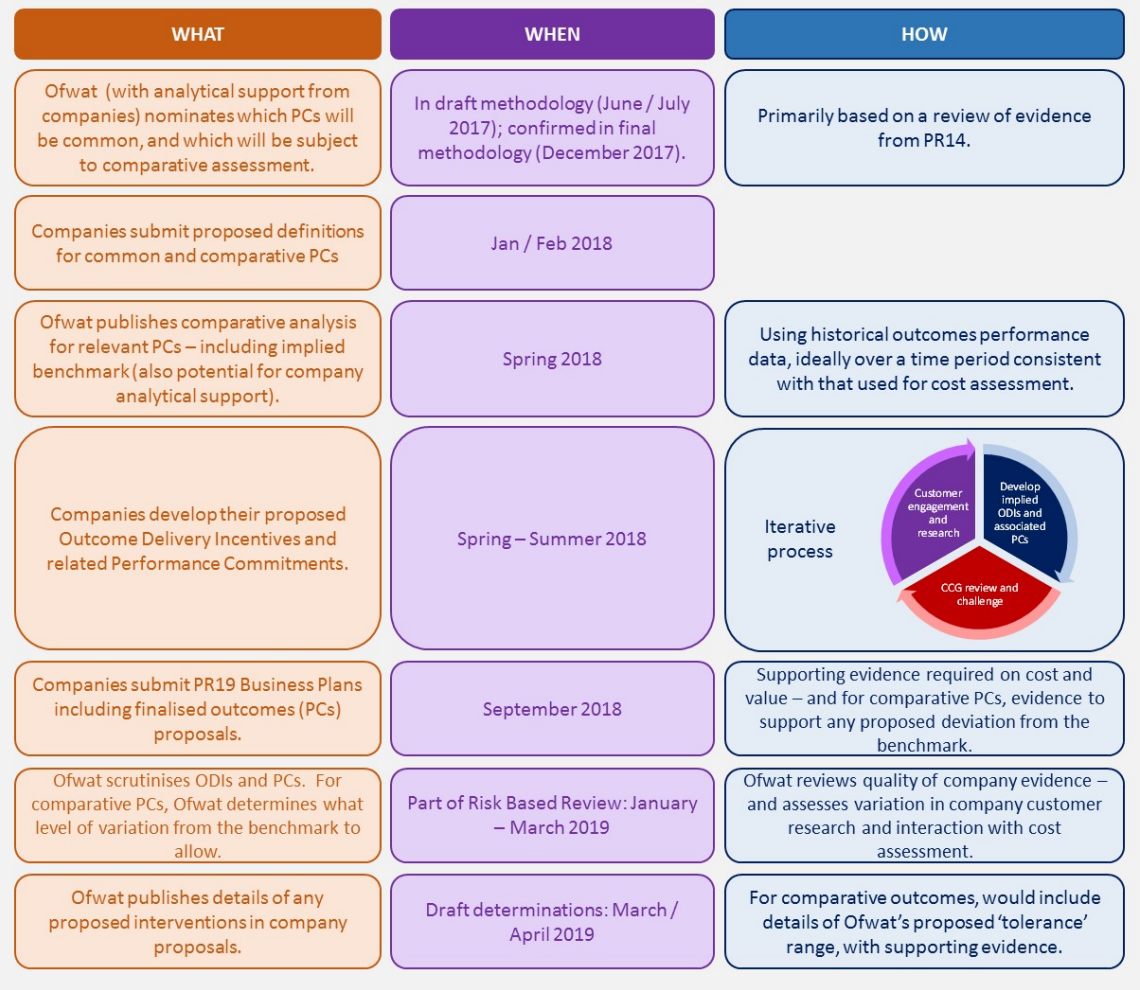
5.1.5. Illustrating how our recommended package would work within the PR19 timeline

The following figure provides an overview of how the process associated with our preferred package could fit within the overall PR19 timetable.

Of course, the figure below provides an illustration of *one possible timeline* and approach for implementing our suggested package. It should not, therefore, be interpreted as the only way of achieving the intended aims.

In addition, whilst the key decisions relating to the control would need to be taken by Ofwat, this does not mean that Ofwat alone would need to take forward all of the required work and analysis to inform those decisions. Indeed, given the relatively tight timeline for PR19, clearly this is an area that would benefit from company input and early engagement – potentially in the form of a working group, for example.

Figure 12: Possible timetable for preferred package



Source: Economic Insight

5.1.6. How comparative information would be used

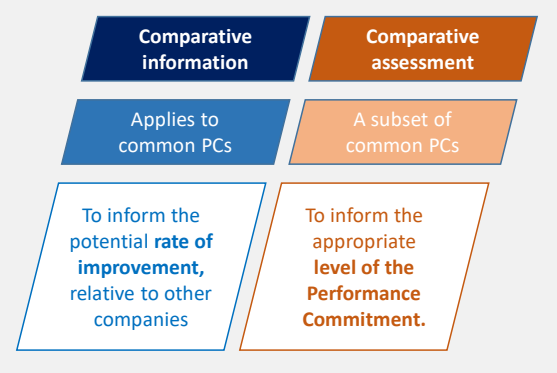
Our proposed approach reflects the fact the we see both 'comparative information' and the results of 'comparative assessment' as being important pieces of information for key stakeholders involved in the development of company plans. Not least, customers and CCGs.

Here, recall that comparative information relates to a subset of performance measures (i.e. data and metrics) that can be used to compare the performance of companies across common PCs. Whereas, comparative assessment refers to undertaking benchmarking across companies to inform the setting of the PC. Following from this, our views are that:

- **Comparative information** for common PCs *not subject to comparative assessment,* is most relevant to inform the potential **rate of improvement** in outcome performance for a company. This is because, as previously explained, if a PC is not deemed suitable for comparative assessment, this would be because one believed the 'economically efficient' level of provision varied materially across companies. Consequently, in these cases, the value of comparative data should not be to compare performance levels. Instead, companies, customers and CCGs might primarily utilise this data to help determine and challenge the potential speed at which companies can make improvements. That is to say, by examining comparative data over time, stakeholders can identify other companies who have made improvements, and then 'challenge' their company (but not set hard benchmarks) to achieve similar gains.
- In contrast, the value of **comparative assessment** (which as explained, should only be used where one expects the economically efficient level of provision to be similar) is that it provides a more meaningful way of challenging the **proposed level** of any PC. For this reason, we see the output of any comparative assessment undertaken by Ofwat itself, as an important piece of information for customers, CCGs and companies.

Figure 13 summarises the above.

Figure 13: Use of comparative information



Source: Economic Insight

By making the outputs of comparative assessment available to CCGs and customers, they will be well placed to challenge any company proposed PC that diverges from the implied 'benchmark' identified by Ofwat.

Specifically, it would be much easier for companies to engage in meaningful dialogue with CCGs around the inherent uncertainty of estimating the costs and value of outcomes. This in turn, would allow CCGs to challenge companies to provide highly robust evidence to justify any proposed deviation from any benchmark. This, in turn, should benefit the RBR process itself.

5.2. How the outcomes framework interacts with cost assessment

5.2.1. Re-cap of cost assessment at PR14 and its relationship with outcomes

Ofwat's approach to cost assessment at PR14 evolved considerably from previous controls. At a high level, it contained the following features.

- **Differential approaches were applied across the different parts of the value chain,** with an average cost to serve approach in retail, compared to more 'traditional' econometric benchmarking approaches in wholesale.
- **The use of a wider range of econometric techniques, which were 'triangulated'.** PR14 saw a greater use of more complex econometric models, such as translog GLS (RE) and COLS, which were *triangulated* in order to get a single estimate for both water and wastewater.
- **An upper quartile efficiency challenge**. In the past, Ofwat benchmarked firms based on an *adjusted* frontier approach. At PR14 Ofwat set the benchmark for wholesale costs at the upper quartile.
- **Consideration of special cost factor claims.** Ofwat considered company representations on any special cost factors not captured by its modelling.

Importantly, there was no explicit link between the above cost assessment and the determination of outcomes incentives. This is true both from a *process* perspective (i.e. the two were not assessed concurrently) and from a *method* perspective (i.e. there was not methodological link between them).

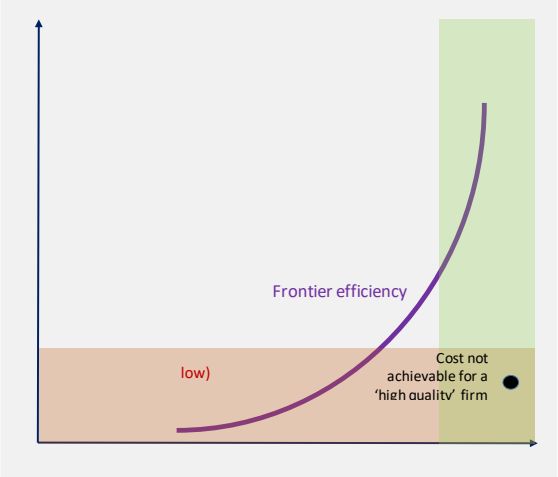
There were, however, implicit links – insomuch that, in reaching its final determinations, Ofwat appraised company plans in their totality. For example, to help decide whether plans appropriately balanced risk and reward across companies, customers and the environment.

5.2.2. Relevant economics and key issues

Referring back to the previous chapter of this report, we identified the fact that competitive markets deliver allocative and technical efficiency. Put another way, a 'frontier' firm is both allocatively and technically efficient. However, regulators face a *measurement problem* in identifying a benchmark for the frontier with precision. In cost assessment, this provides the rationale for either adjusting residuals under a frontier approach, or for using a less demanding benchmark, such as upper quartile. We discuss this measurement issue in section 5.3.

Nevertheless, if we assume that *the benchmark is correctly identified,* the key point is that it should not only provide the appropriate reference point for cost, but also for outcomes. This is because, by definition, the benchmark is consistent with delivering outcomes that are economically efficient (i.e. where the marginal cost of delivering the outcome is equal to the marginal benefit to customers). Put differently, if this benchmark was not the right benchmark for outcomes, nor could it be the right benchmark for cost assessment. This is illustrated in the following figure.

Figure 14: Illustration of the economic frontier for costs and outcomes



Source: Economic Insight

The intuition for the above is simple: that firms face cost quality trade-offs and might choose different cost/quality positions depending on their strategy (for example, Motorola is lower cost than Apple with regard to mobile phones. However, *both could be minimising their economic costs).*

Consistent with the above, we note that a recommendation of the UKWIR report on outcomes was that: *"ideally, service levels and costs would be considered at the same time."*[39] However, we would take a slightly stronger view and suggest that the concurrent assessment of outcomes and costs (where these are assessed comparatively) is, in fact, essential.

This is because, by not 'joining up' the approach to cost assessment and outcomes, the risk of mis-identifying the benchmark is far greater. This could result in companies being either over, or more likely, under-funded – as benchmarks become unachievable over the longer-term. Ultimately, this would harm customers, as companies would not face the appropriate incentives to deliver the outcomes they want.

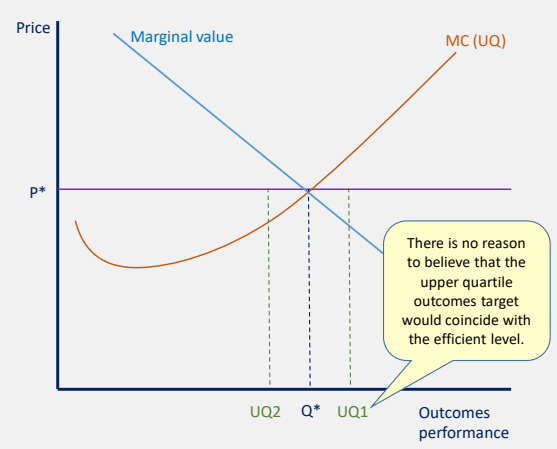
As noted previously, at PR14 Ofwat benchmarked firms to the upper quartile for cost (shown as *MC(UQ)* in Figure 15). Consequently, because the benchmark is both technically and allocatively efficient, this meant that, at PR14, firms were funded for a level of outcome performance consistent with that an 'upper quartile **cost** efficient' firm could deliver (*Q** in Figure 15).

"Whilst the UKWIR report on outcomes recommended that the concurrent assessment of costs and outcomes would be 'ideal', we think it is essential."

When applying comparative assessment to a subset of common PCs, however, Ofwat *independently from cost assessment,* set the performance target for outcomes based on the upper quartile across firms. In practice, this may or may not have been appropriate. What we do know, however, is that there is no *intrinsic* reason as to why the 'outcomes performance' of the hypothetically (upper quartile) efficient firm for cost would happen to coincide with the upper quartile for outcomes across companies.

As a result, the outcomes targets Ofwat set for those comparative PCs could have been below (*UQ2*) or, *more likely* above (*UQ1*) the performance level actually consistent with the funding provided.

Figure 15: illustration of the interaction between cost assessment and outcomes



Source: Economic Insight

[39] *Setting performance commitments and incentives to deliver best value for money.' Frontier Economics, SYSTRA and Tynemarch – UKWIR report 16/RG/07/39 (2016).*

Our analysis concurs with the CMA's view of Ofwat's approach to comparative assessment for common PCs. In its redetermination of Bristol Water's PR14 settlement, the CMA stated: *"for Ofwat to consider that upper quartile performance (historical or otherwise) would match economic levels appeared unlikely to us in general."*[40]

Following from the above, in our view the benchmarking approach to comparative assessment needs to be consistent with cost assessment. That is to say:

- **starting from cost assessment** - for a given cost benchmark (e.g. upper quartile) the associated outcome targets should be consistent with those a hypothetically efficient (upper quartile cost) firm could deliver; or
- **starting from outcomes** – for a given benchmark (e.g. upper quartile) the associated cost assessment should be consistent with that a hypothetically upper quartile firm (in terms of outcomes) could deliver.

Here we should also make two important points of clarification:

- Firstly, the issue we are addressing here is related to, but somewhat distinct from, the previous section of this report. In 5.1 we addressed how one might deal with the inherent *uncertainty* regarding whether the economically efficient performance level was the same, or differed, across companies. To address this, we advocated there being a tolerance zone around any benchmarked PC derived through comparative assessment. Here, however, the issue we are primarily addressing is: *'assuming that the appropriate level of performance is the same (or similar) across companies, how can we ensure it is consistent with the efficient level of allowed costs?'*
- Secondly, the fact that we take the view that a holistic approach to cost assessment and outcomes is required does not mean that, in some cases, companies do not have scope to improve in both dimensions. That is entirely an empirical matter and turns on how efficient they currently are.

5.2.3. Supporting evidence

There is a range of evidence to support the above points – which we set out in more detail in Appendix B. However, to summarise:

- The presence of cost / quality trade-offs in competitive markets is well established in the economics literature. This includes both theory and empirical studies.
- We have undertaken our own analysis of cost / quality performance across industries generally considered to be competitive, which is further consistent with the presence of trade-offs.
- There is a range of precedent in which economic regulators have explicitly recognised the existence of cost / quality trade-offs.

Below we provide some selected examples of the above evidence.

There is a clear trade-off between quality and price in other, competitive, industries

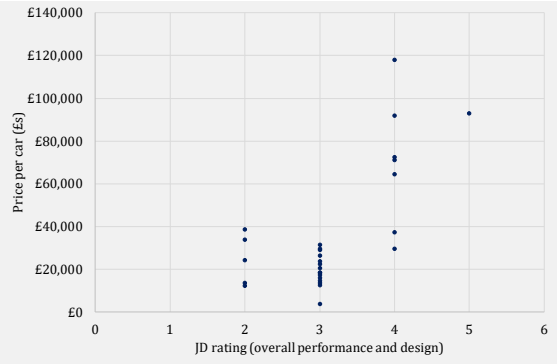
Car manufacturing is generally considered to be a competitive industry. Consumers have good information about price and quality, and are able and willing to substitute across a range of manufacturers to meet their car purchasing needs.

Accordingly, we examined the relationship between the price of new cars and their quality, as measured using the JD Power

[40] *'A reference under section 12(3)(a) of the Water Industry e Act 1991: A Report.' The CMA (2016); page 283.*

survey, which monitors quality performance across four main areas: (i) customer service; (ii) dependability; (iii) quality; and (iv) design. The following scatterplot shows the resultant price / quality performance of car manufacturers.

Figure 16: Relationship between unit price and service level performance across car manufacturers



Source: Economic Insight analysis of JD Power Survey and price data

The above highlights how, in the real world, one tends to see very clear trade-offs between cost and quality in competitive markets. Further details of the method we used to generate these results, and additional analysis for the airline industry (which shows the same pattern) can be found in Appendix B.

The need to concurrently consider cost and outcome performance is established in regulatory precedent

Consistent with the above, there is a range of precedent whereby regulators have acknowledged the above cost / quality trade-offs.

With regard to assessing the efficiency of airports, the CAA has said: *"Higher operating expenditure (opex) per passenger at an airport may reflect many factors including:* ***higher service quality,*** *higher security standards, higher factor costs, lower levels of capital substitution, a wider range of activities being undertaken, different accounting definitions, differences in exchange rates and local prices; economies of scale; as well as underlying efficiency."*[41]

In relation to electricity distribution, Ofgem has considered this issue on a number of occasions. For example, in 2003, Ofgem asked CEPA to review its approach to benchmarking in distribution controls. In its report, CEPA stated that *"there is clearly a trade-off between improving quality and reducing costs."*[42]

CEPA (and Ofgem) ultimately discounted approaches that integrated quality variables into its econometric cost benchmarking models (due to a lack of explanatory power – an issue we discuss further subsequently). However, of relevance to the issues that must be considered for PR19, Ofgem decided to make 'ex-post' adjustments to allowed costs to control for quality. For example, in the 2013 ED-1 control, Ofgem's approach was as follows: *"We have made an adjustment to the upper quartile level of efficiency to take account of instances where DNOs have offered up tighter customer interruptions (CI) and customer minutes lost (CML) targets than our benchmarking methodology has produced."*[43]

With regard to determining the efficiency of BT Openreach in 2012, Ofcom noted: *"In setting an efficiency target consideration needs to be given to the overall regulatory framework, including the time period of operation and* ***quality of service requirements."***[44]

In assessing the efficiency of Royal Mail Group, **Ofcom's approach explicitly**

[41] *'CAA Airport Operating Expenditure Benchmarking Report 2012.' The CAA (2012).*

[42] *'Background to work on assessing efficiency for the 2005 distribution price control review.' CEPA (2003).*

[43] *'RIIO-ED1 business plan expenditure assessment - methodology and results.' Ofgem (2013) page 8.*

[44] *'Charge control review for LLU and WLR services Annexes' Ofcom (2012); page 51.*

accounts for quality performance.[45] Indeed, quality variables are included in the econometric models used to assess Royal Mail's cost efficiency – specifically:

- when assessing the relative of delivery offices, the % of special deliveries delivered on time is controlled for; and
- when assessing the relative efficiency of mail centres, the quality of service in relation to 1st and 2nd class mail is controlled for.

5.2.4. Implications for PR19 approach

The key implication going forward to PR19 is that, where comparative benchmarks for outcomes are imposed in particular, it is important to ensure there is consistency with cost assessment (as otherwise there is a real risk of customer detriment).

This, in turn, raises an obvious question for the PR19 outcomes approach: *how should one best 'join up' cost assessment and outcomes?*

5.2.4.1. Options for achieving consistency

In practice, a consistent approach to cost assessment and outcomes can be delivered by the following one of two *broad approaches:*

- **Exploring alternative options for cost assessment** that 'adjust for' the level of outcome performance delivered by the benchmark.
- **Exploring alternative options to outcome benchmarking** that take account of the 'cost performance' of the benchmark.

Within the above categories, there are a range of *analytical methods* that could be used to deliver an internally consistent approach, these include:

- **Econometric methods.** That is to say within cost assessment, explicitly include outcome related measures and set a benchmark that reflects 'upper quartile' for both cost and outcomes (or vice versa if starting from outcomes).
- **Ex-post modelling adjustment.** Under this approach, rather than include outcome measures in cost modelling (or vice versa) one would instead 'adjust' the implied efficiency targets derived from the cost models to ensure consistency with outcomes. For example, one could calculate the 'cost uplift' associated with upper quartile outcome performance (if that was the benchmark used) and apply this.
- **Special factor regime.** Arguably a variant of the ex-post modelling adjustment above, allows companies to submit claims relating to outcome performance as part of cost assessment. The main downside of this approach would seem to be that it lacks the coherence and consistency of 'across the board' solutions (and so we do not consider it further here).

The first two of the analytical options outlined above are consistent with those identified in the UKWIR report. The only difference being that:

- the UKWIR report 'split out' option 2 (ex-post modelling adjustments) into two categories, based on 'how detailed' the analysis was to inform the adjustment; and
- the UKWIR report did not raise the prospect of a special factor type approach.

In relation to the options it considered, the UKWIR report also stated: *"the key question for the water sector is whether the cost allowance should be adjusted on the basis of a small number of industry measures or*

[45] *'Econometric benchmarking in the UK postal sector: Final Report.' Deloitte report for Ofcom (2016).*

whether that would imply too much emphasis is put on these measures."[46]

We do not agree with this sentiment, because it seems to give rise to a tension between, on one hand, saying that the outcomes are sufficiently important to customers that it is right to include them in an outcomes framework and to set comparative benchmarks using them (with financial implications); but on the other hand, might not be sufficiently important to require an automatic cost adjustment. Nonetheless, the options identified in the UKWIR report align closely to those we have identified and considered here.

When considering the relative merits of the potential options, we think an important point of principle is that any approach should not be expected to result in:

- more or less accurate assessments of efficiency for any one company compared to others; and / or
- differences in the implied cost challenge across firms that are not related to any underlying economic characteristic.

In turn, the above points to an approach where:

- a *consistent analytical method* is applied across companies; and
- as a point of principle, any adjustments to allowed costs are *applied to all companies* (although the size of adjustments could clearly vary across firms and might even be 'negative' in some cases).

Having identified the analytical methods by which the consistency of cost assessment and outcomes could be achieved, in the following section we build further on the UKWIR report by providing an assessment of their feasibility.

5.2.4.2. Testing the feasibility of the options

In this section, we provide our assessment of the econometric and ex-post modelling adjustment options for achieving consistency in the approach between outcomes and cost assessment. Note, our aim here is not (within any particular approach) to identify a single 'right' method in any detail (which would be premature at this stage). Rather, our analysis is intended to help identify which approach is most feasible – so that further work could then be taken forward to help identify a more detailed method, consistent with that approach.

The feasibility of econometric options

To examine the feasibility of using econometric methods to achieve consistency, we replicated two of the water econometric models used by Ofwat to set the cost benchmark (WM5 and WM6) at PR14 and two of the wastewater models (SW9 and SW10). In doing so, we added into the model the comparative outcome measures relevant to each area – that is:

- for ***water*** we included – (a) the average number of water quality contacts per '000 of the population, (b) the average length of supply interruptions per property, and (c) the water drinking quality measure (mean zonal compliance %); and
- for ***wastewater*** we included – (a) pollution incidents per km of sewer and (b) properties subject to sewer flooding per '000.[47]

Overall, we found that the adjusted models did not work well. Specifically, as illustrated by the following tables, the coefficients on

[46] *Setting performance commitments and incentives to deliver best value for money.' Frontier Economics, SYSTRA and Tynemarch – UKWIR report 16/RG/07/39 (2016); page 47.*

[47] *Specifically, we included the (natural logarithm) of the three-year average used by Ofwat to set the PCs.*

the quality variables often had the wrong sign and/or were statistically insignificant in at least one of the models.

Table 6: Water econometric models including comparative quality measures

	Correct sign?	Statistically significant?
Mean zonal compliance	WM5: Yes WM6: No	WM5: No WM6: No
Water quality contacts	WM5: Yes WM6: Yes	WM5: Yes WM6: No
Supply interruptions	WM5: No WM6: No	WM5: Yes WM6: Yes

Source: Economic Insight analysis of PR14 cost assessment data

Table 7: Wastewater econometric models including comparative quality measures

	Correct sign?	Statistically significant?
Pollution incidents	SW9: Yes SW10: Yes	SW9: No SW10: No
Internal sewer flooding	SW9: No SW10: No	SW9: Yes SW10: Yes

Source: Economic Insight analysis of PR14 cost assessment data

Moreover, the inclusion of quality measures led to large swings in the size and significance of other variables in the models.

When considering why the econometric approach has not yielded robust results, there are a number of potential explanations, which require careful consideration.

- that there is **no relationship between cost and outcomes;**
- that **the 'measures'** of cost (efficiency) and / or the measures of outcomes **are 'wrong'** (which could include: that the cost assessment models omit certain cost drivers – including those relating to outcome performance; and / or that the definitions of outcomes measures vary across companies); and / or
- **modelling related issues**, such a multicollinearity.

Of course, the above explanations are not mutually exclusive; and a range of factors may be at play. However, for the reasons already outlined and evidenced previously, the first explanation (that there is no cost / quality relationship) seems unlikely to be the main factor.

If the results reflect problems in the cost measures or outcomes measures themselves, this of course raises more fundamental questions for PR19 (which would apply across whichever method one selected here in relation to ensuring consistency of outcomes and cost assessment). For practical reasons, therefore, we should set this aside for the purpose of considering the 'feasibility' of the various approaches.

The feasibility of ex-post modelling options

In the event that outcomes performance cannot be adequately reflected directly in cost assessment through econometrics (as above) ex-post modelling adjustments might be used instead.

In line with the above, the ex-post modelling options would broadly involve either:

- first, adjusting the PCs set via comparative assessment such that they are consistent with the performance level attained by the efficient (e.g. upper quartile) firm on the cost assessment analysis; or
- second, adjusting the cost allowance, such that firms are funded for the level of performance they deliver relative to the efficient firm in relation to outcomes performance; or
- third, taking account of firms' relative quality performance levels when deciding on what cost efficiency challenge to set them.

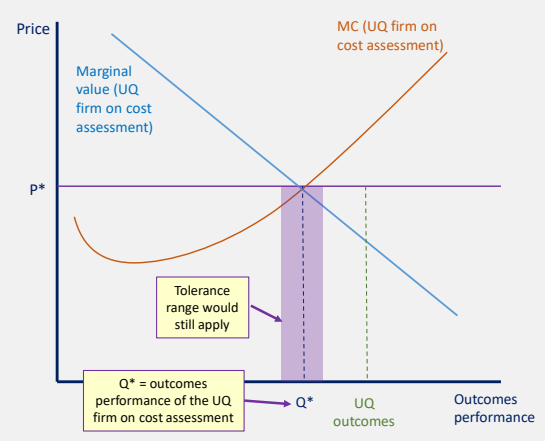
Ex-post adjustment method 1: adjusting the benchmark level of PC

The first adjustment could involve:

- identifying the upper quartile firm in the *cost assessment analysis*[48];
- identifying the performance levels achieved by that firm in each outcome area of relevance (i.e. common outcomes subject to comparative analysis); and
- setting the benchmark PCs for other firms at these levels.

This is illustrated in the following figure.

Figure 17: Illustration of how consistency can be achieved by adjusting the PC



Source: Economic Insight

The key features of this approach are that:

» The benchmark for common outcomes subject to comparative assessment becomes *the outcome performance of the upper quartile firm on cost assessment.* This may, or may not, also coincide with the upper quartile for outcomes performance. However, there is no in principle reason why one would expect it to coincide. In the above figure, for illustrative purposes only, the implied benchmark (Q*) for outcomes is below the UQ level of outcome performance (UQ outcomes). However, in some cases the implied benchmark could be *above* the upper quartile.

» For reasons explained in Section 5.1 of this report, **the 'tolerance band' around the outcomes benchmark would still apply.** This is to reflect the fact that we cannot be certain that differences in the observed level of outcome performance across firms just represent 'inefficiency' (due to both imperfections in cost assessment, and the measurement of customer value). Moreover, as noted previously, one would *never* expect the economically efficient level of outcome performance to be exactly the same across companies.

The main advantage of this approach is that it is straightforward to implement, and is highly transparent. The main area of subjectivity is in determining which firm (or firms) close to the benchmark (UQ) on cost assessment provides the most appropriate reference point for setting the benchmark on outcomes.

The downside of this approach is that it assumes (for any outcome to which it is applied) that the benchmark firm(s) for cost assessment cannot improve its outcomes performance (due to the cost / quality trade off issue). In practice, this may, or may not, be true. Therefore, Ofwat might be concerned that such an approach could result in an unduly lenient PC in some cases. On the other hand, for reasons already set out, making no adjustment at all is also demonstrably not appropriate.

One way of addressing this would be to apply the adjustments only to outcomes where Ofwat felt the evidence of a relationship with costs was strongest. This, in turn, would ideally require a more detailed analysis of the incremental costs of delivering outcomes, which we discuss

[48] *Or in practical terms, a firm or firms 'close to' the upper quartile position on cost assessment.*

further below. However, if that was felt impractical for PR19, Ofwat could instead reach a view as to which outcomes the adjustment might be appropriate for based more on 'judgement' and industry views.

Ex-post adjustment method 2: adjusting allowed costs

Rather than adjust the benchmark PC, the second approach involves mechanistically adjusting the level of allowed cost. This would involve the following steps:

- identifying the upper quartile (or whichever benchmark Ofwat intends to use) firm in relation to *outcomes* using comparative analysis (for the relevant common PCs);
- calculating the differences in the outcome performance for the upper quartile firm (for outcomes) and all other firms;
- multiplying these differences by the incremental cost of increasing (or decreasing) firms' level of outcomes performance in order to meet the benchmark; and
- finally, making upward or downward adjustments to the cost allowance to reflect these differences.

From a computational perspective, this approach is relatively straightforward. Specifically, for any firm, the ex-post adjustment to allowed costs would be as follows:

$$Ca = [OPuq_{,} - OPa] * INCa$$

Where:

- *Ca* = the adjustment to allowed costs for firm *a*
- *OPuq* = the outcome performance of the upper quartile firm for outcomes
- *OPa* = the outcome performance of firm *a*
- *INCa* = the incremental (efficient) cost for firm *a* of increasing its outcome performance to *OPuq*

The challenge with this approach is to accurately identify the relevant incremental costs. That is to say, the extent to which the (efficient) costs of provision genuinely vary with the level of outcome performance.

In order to address this, a combination of 'top-down' and 'bottom-up' approaches to cost modelling could be deployed.

- By **'top-down'** we mean statistical type approaches that seek to identify relationships between firms' costs and outcomes performance. Here, the fact that our econometric approach to achieving consistency between outcomes and costs failed to identify relationships might indicate that these methods will be challenging. However, it might be that a more granular top-down approach would be more fruitful. Here, this could involve specifically identifying just the sub-set of costs relevant to the outcome(s) in question (rather than wholesale totex) and then seeking to develop regression analysis that controls for the drivers relevant just to those costs, *including historical outcome performance.*
- By **'bottom-up'** we mean developing a detailed mapping of how the relevant outcomes might interact with cost drivers, and so in turn, costs (again most likely focusing on just a sub-set of wholesale totex). At the core of this is a need to clearly identify the processes and steps firms must implement in order to deliver certain outcome performance levels, and then putting detailed costings against those processes. This approach would most likely require operational and engineering expertise.

Regarding the above, it is important to recognise that precisely identifying the (efficient) incremental costs of improving outcomes performance is unlikely to be achievable. Therefore, under any method, some subjectivity will remain.

Notwithstanding this, the goal under this option would be to draw on the evidence (ideally both top down and bottom up) in order to derive cost adjustment factors that can be applied: (i) transparently; (ii) consistently across the companies; and (iii) mechanically, as part of the price control process.

Ex-post adjustment method 3: subjective consideration of outcomes performance when setting efficiency scores

Another approach would be to make adjustments to the cost efficiency challenges in a less mechanistic way, to recognise that some companies perform better than others in terms of outcomes. For example, where the cost assessment analyses suggest a range of efficiency scores for companies (across which Ofwat could triangulate) the regulator could pick the less challenging end of efficiency scores for companies that have relatively high quality performance levels.

Whilst this approach is arguably the most practical of the three options identified, it runs the greatest risk of there being 'inconsistencies' across companies. Consequently, it would seem less well suited to promoting efficiency.

Summary assessment of feasibility

In principle, one might imagine that both the econometric analysis and ex-post modelling analysis are "as feasible as each other". This is because, in broad terms, the data needed to implement them is similar.

However, our analysis suggests that econometric analysis may not lead to usable results (certainly at the wholesale totex level). This, points to pursuing an ex-post modelling adjustment approach. Of the three options for making such adjustments, we favour the first two (mechanistic adjustments to either the PC or cost allowance for common PCs subject to comparative assessment) over the third, more subjective, method.

Both mechanistic options would ideally be based on more detailed analysis of the interaction between outcomes and costs. If such analysis was feasible within the PR19 timetable and resource constraints, we would be relatively indifferent between methods 1 and 2. If such analysis is not achievable, however, we would on balance favour the approach of adjusting the benchmark PC (i.e. method 1). Here Ofwat could deal with the uncertainty as to the extent to which outcomes impact cost by only adjusting *some* comparative PCs where, in its judgment (and based on industry views) the relationship with cost was likely to be strongest.

Whilst clearly no method is 'perfect', it is important to recall that we have previously established that a failure to assess costs and outcomes in a consistent manner is inappropriate as a matter of principle. Consequently, we consider developing an approach to address this is an essential element of building on the successes of PR14.

> *"Our analysis suggests that econometric analysis may not lead to usable results. This points to an ex-post modelling adjustment approach."*

5.3. Determining the appropriate measure of the benchmark

5.3.1. Defining the benchmark measure

The choice of *benchmark measure* relates to whether any targeted level of performance for common PCs (subject to comparative assessment) should continue to be based on upper quartile performance, or something more or less stringent. Specifically, in the recent workshop on outcomes, Ofwat suggested three alternatives to an upper quartile challenge, which we consider here:

- upper quintile;
- average of the best three performers; or
- frontier.[49]

5.3.2. Relevant economics and key issues

The purpose of using comparative assessment to set performance targets (for a subset of common PCs) is to encourage companies to deliver frontier levels of performance (see Chapter 4 of this report). However, regulators face a practical challenge when setting benchmarks, in that it is not straightforward to identify the frontier precisely.

This challenge arises for a number of measurement-related reasons, including:

- first, the *measures used may be imperfect* proxies for the underlying outcomes of interest;
- second, the measures may be subject to *measurement error* e.g. due to sampling limitations;
- third, outcomes may be *more achievable for some companies than for others*, due to factors outside of company control; and, relatedly
- fourth, the *measures may be affected by 'one-off' or unusual events* from year-to-year.

The implication is that the level of performance achieved by the 'best company' according to the outcome measure used may not, in fact, be a good proxy for the *frontier* level of performance.

This problem is well-known and motivates why regulators, in the context of cost assessment, tend not to apply a 'pure' frontier approach when setting benchmarks (i.e. they do not simply select the 'most efficient' company as identified by their benchmarking analysis). Rather, regulators instead: (i) apply adjustments to the identified frontier, such as reducing the residuals in regression analysis assumed to relate to inefficiency; (ii) discount certain firms altogether from representing the frontier, typically based on judgement about the extent to which company specific factors make said firm an unsuitable benchmark; or (iii) set the benchmark based on upper quartile, or some other measure, of performance, rather than the 'most efficient' company.

This is summarised by the ORR as follows:

> *"COLS models assume that all noise is inefficiency, which is implausible. Ofwat, in their 2007-2008 relative efficiency assessments adjusted COLS residuals for water by 10% and for sewerage by 20%. In contrast Ofgem have benchmarked against the upper quartile for the most efficient firms and the upper third or average for less efficient firms. In keeping with the view that there is some noise in the data that the COLS approach fails to account for, we have decided to apply what we would view to be a conservative 25% noise adjustment – our assumption is that, on average, 25% of an IM"s deviation from the frontier can be attributed to noise."*[50]

[49] *'Outcomes – Water 2020 stakeholder workshop.' Ofwat (14 June 2016).*

[50] *'PR13 Efficiency Benchmarking of Network Rail using LICB.' ORR (2013).*

The key points to highlight are as follows:

- Conceptually, the benchmark is intended to represent / encourage frontier levels of performance.
- However, because the frontier cannot be precisely identified, it is widely recognised that picking the 'most efficient' firm is inappropriate. Accordingly, whether one sets the benchmark based on the frontier, but with adjusted residuals, upper quartile, or so on – ultimately **these are all different ways of addressing the measurement problem.**[51]
- Consequently, the choice between one benchmark and another essentially turns on **how challenging** the regulator thinks it is appropriate to be, given that it wants to: (i) encourage **frontier levels of performance**; whilst (ii) **not setting unachievable challenges** (which could dampen the incentive power of the cost assessment regime).

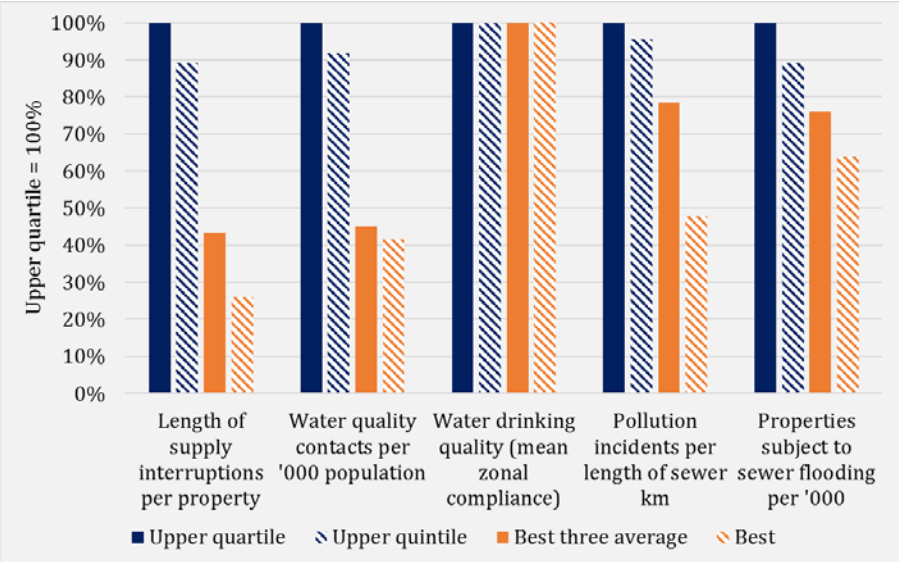
5.3.3. Implications for PR19 approach

To help evaluate the implications of the above for the PR19 approach, we examined the extent to which using different benchmarks could result in different PCs, compared to continuing to use the upper quartile benchmark. Without such differences, arguably the choice of benchmark would not be a significant decision at PR19.

The figure below presents the results of our analysis. It shows how *'different'* the PCs would be based on the different benchmark measures compared to the upper quartile measure used at PR14 for the comparative ODIs (excluding leakage).

So, for example, the chart shows that in the case of supply interruptions, the best performing company has a performance level that is 26% of the performance level of the upper quartile company (i.e. the upper quartile performance is 10.4 minutes per property, whereas the best performance is 2.7 minutes per property).

Figure 18: Comparisons of different benchmark measures



Source: Economic Insight analysis of Ofwat 'upper quartile comparative assessments' data.

[51] *Where by measurement problem we are referring to a range of issues that might result in the benchmark being misidentified. This could include: omitted variables (e.g. not controlling for factors that genuinely drive cost differences, such as historical investment patterns or other regional factors); inconsistencies in how variables are defined across companies; inconsistencies in how data is recorded across companies – and so on.*

The chart shows the following:

- first, a move to either a *best* or *average of best three* benchmark would, in all cases except for water drinking quality, **require even the upper quartile company to make substantial improvements**;
- second, though the upper quintile benchmark is much closer to the upper quartile benchmark, **it would still require the upper quartile company to improve its performance by 5-10%;** and
- third, the size of performance improvement implied by the different benchmarks **varies from measure to measure.**

The implications for the PR19 approach are as follows.

- **The 'best' or 'average of best three' benchmarks represent very demanding challenges, even for upper quartile firms.** Therefore, to adopt such benchmarks, one would have to be very confident that the implied challenges are achievable and not an artefact of the measurement issues outlined above. The fact that there is such a large 'jump' between the upper quartile performance level and these benchmarks is cause for concern in this regard.
- **The improvement required to meet the different benchmarks varies materially from one outcome to another.** Therefore, there may be a need to evaluate whether a particular benchmark is appropriate on an outcome-by-outcome basis – one size might not fit all. However, one would need to be mindful of adding undue complexity – and further, this issue needs to be considered in the context of whether it is more appropriate to have 'individual' or 'aggregate' outcomes benchmarks, as we discuss subsequently.
- **The choice between the upper quartile and upper quintile benchmark is arguably more finely balanced.** Subject to data being available, one way of choosing between them would be to evaluate how easily or otherwise companies achieved the upper quartile performance level during PR14.

5.4. Whether outcomes benchmarks should be set on an individual or aggregate basis

5.4.1. Defining individual and aggregate approaches

At PR14, comparative benchmarks (for a subset of common PCs) were set individually, so that a separate PC, with an associated penalty or reward where applicable, was set for each company for its performance *against each individual* area. Alternatively, a benchmark could be set on an *aggregate* basis, with reference to firms' overall performance across a number of PCs.

The main practical difference is that an individual approach would tend to be more demanding for firms compared to an aggregate approach. Simply, under an individual approach, a firm will bear performance risk in relation to each area separately (and so under-performance in any one area would result in a penalty). Whereas, under an aggregate approach, under-performance in one area might be offset by over-performance in another; and any penalty or reward depends only on the overall performance.

Following from the above, there are three key questions to consider:

- What issues determine whether it is appropriate to use an aggregate or individual benchmark?
- What evidence is there to inform the choice of aggregate or individual benchmarks?
- How can an aggregate benchmark be determined in practice?

5.4.2. What determines whether aggregate or individual benchmarks are appropriate?

In simple terms, if there are trade-offs across performance in outcome areas, an aggregate approach is more likely to promote allocative efficiency, as it provides firms with the flexibility to focus their efforts accordingly. On the other hand, if there are no such trade-offs, or the trade-offs are limited, a more focused 'individual' approach may drive stronger technical efficiency incentives. Therefore, at the heart of this issue, is whether such trade-offs exist. There is both a customer, and cost structure, dimension to his.

5.4.2.1. Consumer preferences

As we have previously explained, the 'spirit' of outcome incentives is to encourage delivery up to the economically efficient level. We therefore explained that comparative approaches make most sense where one believes the economically efficient level of provision might be 'the same' or 'similar' across companies.

Following from this, **an aggregate approach would make the most sense where:**

- One believed that customer valuations of 'overall' outcome performance were relatively similar across companies, but that, at a more detailed level, **customer preferences and valuations of specific outcome measures within that overall measure varied across companies.**
- Related to the above, there was generally **greater uncertainty regarding the similarity of customer valuations** at a more detailed level, relative to an 'aggregate' level comparison. Consequently, one would be 'more confident' that at aggregate benchmark would create the appropriate incentives, relative to more individual incentives.

In the above cases, the advantage of an aggregate approach is that it would focus the technical efficiency challenge on a broader metric, where one might be more sure that the economically efficient level was similar across firms.

On the other hand, an individual approach would be preferable where one could establish that customer valuations were

similar / the same across companies for each individual area. This is because, in these circumstances, an individual approach to benchmarking outcomes would give firms a stronger technical efficiency incentive (with no allocative efficiency downside).

We highlighted previously that a key challenge at PR14 was there generally being a lack of process or data with which to objectively determine where, and when, customer value genuinely varied across firms. This same challenge, therefore, is central to the choice between individual and aggregate approaches.

5.4.2.2. Cost structures

Just as customer values of certain outcomes (either individually or in totality) might vary across firms, so too might the economically efficient costs of achieving certain outcomes. Consequently, even if customer valuations were consistent across firms for *each individual outcome,* differences in efficient costs across firms might also mean that there are 'trade-offs' in outcome performance across individual outcome areas.

To see this, suppose that two firms' customers have identical preferences, but the first firm finds it costlier to improve outcome B and the second finds it costlier to improve outcome A. Again, there are penalties attached to underperformance on both individual and aggregate measures. In this case, under an *individual measure,* both firm A and firm B would have incentives to avoid low performance on *both* measures. Given the differences in their costs, however, it may be that the choices that would maximise value to consumers actually involve the first firm concentrating on outcome A; and the second firm concentrating on outcome B.

This suggests that when there is significant variation in firms' (efficient) costs of achieving particular outcomes, it could be beneficial to use an aggregate measure. This would give firms flexibility in pursuing the outcomes that they can achieve at lowest cost, thereby maximising benefits for consumers. Here, following from our discussion in the previous sections of this report, this is a matter of degree. That is to say, if the economically efficient costs of provision *varied materially,* one might not apply comparative assessment to the common PC in the first place. In this case there would be no benchmark at all, and so the debate between 'individual' and 'aggregate' approaches becomes redundant.

In addition to the above, there may be interdependencies between delivering outcomes within any individual firm. This also could result in trade-offs, insomuch that incurring cost or investing effort in any one outcome area might directly or indirectly adversely affect performance in another outcome area.

5.4.2.3. Bringing together costs and valuations

In the following table, we summarise the likely implications of differences and similarities in customer valuation and the (efficient) costs of delivering outcome performance. In doing so, we highlight where individual, or aggregate, benchmarks are likely to be most appropriate.

Table 8: Interaction of consumer valuations of outcomes and costs to firms of achieving them: implications for individual or aggregate benchmarks

	Consumers' valuations of outcomes are the same (similar) for individual outcome areas, across companies	Consumers' valuations of outcomes differ for individual outcomes, across companies
Firms' efficient cost of delivering outcomes is the same for individual outcome areas	Individual likely to be superior	Aggregate likely to be superior
Firms' efficient cost of delivering outcomes varies for individual outcome areas	Individual may be superior if the regulator can measure and control for differences in efficient costs; otherwise aggregate superior	Aggregate likely to be superior

Source: Economic Insight

5.4.3. Rewards and penalties under individual and aggregate benchmarks

When regulators determine the overall incentives associated with their outcomes framework, they typically take a view as to what the total amount of 'value at risk' should be. Accordingly, as an aggregate approach provides firms with additional flexibility, the implication might be that:

- were Ofwat to adopt an aggregate approach, it might set a 'more demanding' benchmark; whereas
- were Ofwat to adopt an individual approach, it might set a 'less demanding' benchmark.

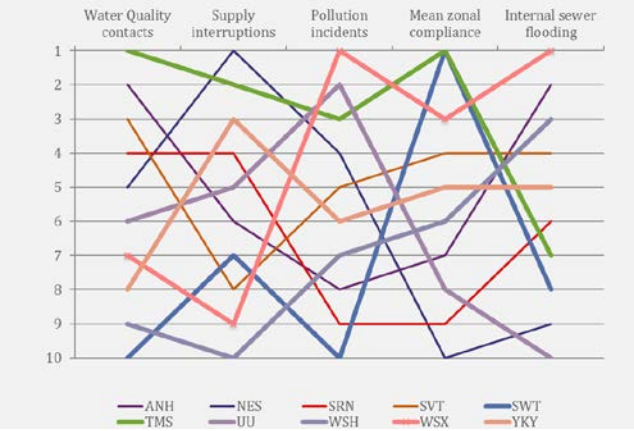
5.4.4. Evidence on the trade-offs in outcome performance

Given the fundamental issue of there being a lack of good evidence to determine whether, and to what extent, customer value varies across the companies – it is hard to reach any definitive conclusions as to whether individual or aggregate benchmarks are most appropriate in the water sector at PR19. However, there is some evidence that can be used to inform this. This evidence tends to suggest that there are trade-offs across outcome areas.

The details of this are set out in Appendix D. However, in summary we find that:

» In relation to PR14 outcomes measures, the same company was only the 'frontier' on two occasions and that there is considerable variation in the ranking of individual firms across the comparative outcome areas (see following figure).

Figure 19: WaSC's rankings across comparative outcome areas



Source: Economic Insight analysis of Ofwat 'upper quartile comparative assessments' data.

» In markets generally considered to be competitive, even highly performing companies (in terms of overall outcomes) have varying performance when one

analyses outcome performance at a more granular level. For example, for both car manufacturers and airlines, we find that the 'frontier' at an aggregate level still delivers sub-frontier performance in some individual categories.

5.4.5. Determining an aggregate benchmark in practice

If an aggregate outcomes benchmark was considered appropriate for water, how should such a benchmark be constructed? There are two main issues that will be encountered in the construction of an aggregate benchmark across any common PCs subject to comparative assessment.

- First, given the range of concepts that the performance commitments measure, ranging from pollution incidents to numbers of water quality contacts, how can these be made 'comparable' to enable their use in the benchmark?
- Second, in combining performance across the outcomes, what weight should one attach to each particular outcome?

With respect to the first question, we note that the key point of interest is generally how far away a company is from 'frontier' (i.e. as proxied by upper quartile at PR14) performance. The challenge in combining measures is to remove the 'units' with which different distances from the frontier are calculated. This can be done by calculating the number of standard deviations that firms are from the frontier (we have undertaken an analysis to illustrate what this would imply, set out in Appendix D).

To decide on the appropriate weight to attach to particular outcomes, we note that, irrespective of the benchmark, firms will have an incentive to minimise costs. The benchmark weighting should, therefore, concentrate on the value that consumers place on particular outcomes. Here, the most obvious approach would be to weight by the 'average' value customer attach to the outcomes across companies (which by definition should be similar, given that this is a necessary condition under our 'decision tree' for determining when comparative assessment should be applied in the first place).

5.4.6. Implications for PR19

At this stage, there is insufficient evidence for us to form a strong view on the relative merits of individual, or aggregate, approaches to benchmarking. The lack of consistency in rankings across firms in individual areas, coupled with evidence from other markets, points toward considering an aggregated approach. However, better information on the extent of trade-offs across outcome areas – driven both by variation in customer valuations and company costs – would be critical to inform any final decision.

> *"The evidence points toward considering an aggregated approach. However, better information… would be critical to inform any final decision."*

5.5. The pros and cons of setting benchmarks on a static, or dynamic, basis

5.5.1. Defining 'static' and 'dynamic' benchmarks

In this section we set out working definitions of static and dynamic benchmarks. Here, recall that benchmarks of any kind would only apply to a subset of *common PCs* for which *comparative assessment* is deemed to be appropriate.

For the purpose of this report, we adopt the following definitions.

- A ***static benchmark*** is a benchmark that results in annual PCs being set at the start of a price control period and remaining unchanged during it. For clarity, a static benchmark could involve different annual PCs in each year of the price control – but the *levels* are determined at the start of the price control period.
- A ***dynamic benchmark*** is a benchmark that results in annual PCs being set at the start of a price control period, but that can change during the price control period.

These definitions are consistent with those previously used by Ofwat,[52] UKWIR[53] and others.

Ofwat's definition (referenced above) specifically mentions improvements made '*during the price control period*' – this is important. This is because the distinction between a static and dynamic benchmark is blurred across price control periods. That is, one could reasonably view a static benchmark ***during*** *a price control period* as a dynamic benchmark ***across*** *price control periods*. This matters, given that Ofwat is also considering whether to apply PCs across price control periods.

5.5.2. The different types of static and dynamic benchmarks

The theoretical and practical pros and cons of static and dynamic benchmarks are likely to vary, depending on their exact form; and so, for PR19, it will be important to refine the high-level definitions set out above.

Ofwat provided a useful starting point in its June 2016 industry workshop. Specifically, it noted that there are different types of static and dynamic benchmarks, namely:

- static historical (i.e. the PR14 approach);
- static forecast;
- dynamic pure; and
- dynamic ratchet.[54]

With the above in mind, below we set out the main features that define different types of static and dynamic benchmarks.

5.5.2.1. Features of static benchmarks

The difference between a *static historical* and *static forecast* is that the former is based on *known* performance at the time of setting the price control, whereas the latter is based on *expected* performance. Another variation is to use a *combination* of both historic and expected performance.

The key issues to consider relevant to the above are:

- first, when using *historic* information, how best to strike the right balance between using the latest evidence available, versus attaching 'too much' weight to one or two years, which could be 'unusual' or 'one-offs'; and

[52] *"Both the approaches to setting UQ levels for wholesale costs and PCs subject to comparative analysis are static and based on historical data and do not dynamically adjust during the price control period to take into account improvements over time." 'Final price control determination notice: policy chapter A2 – outcomes', Ofwat (12 December 2014), page 36.*

[53] *"Dynamic PCs are ones in which the target evolves over time, and are usually dependent on performance in some way. For example, the target for a given year could be set at the upper quartile of performance in the previous year." 'Setting Performance Commitments and Incentives to Deliver Best Value for Money', UKWIR (2016), page 18.*

[54] *'Outcomes – Water 2020 stakeholder workshop.' Ofwat (14 June 2016).*

- second, when using *forecast* information, how best to determine the accuracy of the forecasts (which could itself rely on using historic information).

5.5.2.2. Features of dynamic benchmarks

Dynamic benchmarks can vary in the four ways set out in the table below.[55]

Table 9: Choice of benchmark

1. Individual or comparative revision
The extent to which changes in an individual company's performance commitment is determined only by *changes in its own absolute performance* (relative to the PC it was set at the start of the price control period) or *changes in its performance relative to other companies.*
2. Speed of revision
How quickly new performance information results in changes to a company's PC. For example, the regulator could decide to adjust a PC *as soon as new performance information becomes available*, or wait until it is *validated by another year of performance information.*
3. Revision weight
How much new performance information results in changes to a company's PC. For example, the regulator could decide to set a new PC *based exclusively on the new upper quartile* (UQ) performance level or, say, *the simple average of the old upper quartile and new upper quartile.*
4. Revision symmetry
Whether revisions to the PC are *only allowed in one direction or both* e.g. the PC is only changed if UQ performance is better, not worse (asymmetric).

Source: Economic Insight

Following from the above, and for example, one 'special case' of a dynamic benchmark would involve adjusting the PR14 approach, so that companies get a new UQ PC as *soon as Ofwat knows that UQ performance has changed.* This would involve:

- comparative revision;
- immediate revision;
- 100% weight attached to the new information, 0% weight attached to historic information; and
- symmetric revision.

We discuss the pros and cons of such an approach in the next sub-section, but here we note that it would only make sense if, at a minimum, one could be confident that:

- the observed change in UQ performance **is an artefact of a real and persistent change in company performance**, not due to a one-off event outside of company control; and
- **companies had the ability and incentive to make swift changes** to help meet the new PC.

We are not aware of specific definitions of *dynamic pure* or *dynamic ratchet* benchmarks and do not seek to do so here (we are cautious about creating confusion). However, one possibility is that a pure benchmark involves immediate revision, whereas a ratchet benchmark introduces a delay. We note that *any* dynamic approach would seem to create a 'ratchet' effect to some degree, by virtue of the fact that they modify the 'target' PC based on changes in actual performance (and so, should performance improve *during a control period,* the PC would become more demanding).

[55] *For clarity, this section addresses the how the benchmark changes during the price control period, it does not consider whether penalties and rewards should be based on average performance during the price control period or performance in each year. Whether this choice matters depends on whether there are symmetric rewards and penalties.*

5.5.3. Relevant economics and key issues

In this section we set out the relevant economics and key issues associated with the use of static and dynamic benchmarks – and, in doing so, indicate situations in which one type might be preferable to another. In summary, we consider that there are three main issues:

- first, the incentives the benchmarks create for companies to deliver ***technical, dynamic and allocative efficiencies***;
- second, the extent to which the outcome (or outcome measures) are ***controllable***; and
- third, the costs and benefits of creating ***uncertainty*** within a price control period.

5.5.3.1. Incentives for technical, dynamic and allocative efficiency

The first economics issue is the extent to which different types of benchmark create incentives for companies to deliver technical, dynamic and allocative efficiencies, as defined in Chapter 4 of this report.

With this in mind, there is significant (and helpful) overlap between the economics of static and dynamic benchmarks and the economics of price control duration. To illustrate this, under a static benchmark, a company that out-performs the PC set at the start of the price control period would earn rewards for the duration of the price control, but it would not under a dynamic benchmark.

The implication is, in line with the economics of price control duration, that **a static benchmark is more likely to promote technical efficiency, whereas a dynamic benchmark is more likely to promote allocative efficiency.**

The choice between using a static benchmark or dynamic benchmark is primarily a choice between whether to promote technical efficiency or allocative efficiency.

- ***Technical and dynamic efficiency.*** It is well established that the *longer* a price control period is, the *stronger* the incentive for regulated companies to improve technical efficiency. The reason for this is that companies benefit from any out-performance (and suffer from any under-performance) for longer before they are 'clawed back' by the regulator. Similar arguments relate to the incentives for improving dynamic efficiency. **Dynamic benchmarks could, accordingly, reduce the benefit of out-performance and thus the incentive to out-perform in the first place.**
- ***Allocative efficiency.*** To provide an incentive to improve technical efficiency, a static benchmark 'disconnects' prices from costs for a period of time; and therefore, creates some allocative inefficiency. In contrast, **the benefit of a dynamic benchmark is that it could help prices 'keep in line' with costs,** and reduce such distortions.

Therefore, **the choice between using a static benchmark or dynamic benchmark is primarily a choice between whether to promote technical efficiency or allocative efficiency.**

This choice could point to different approaches for different outcomes, as illustrated by the following table. For example, in an area where it is likely that

consumer tastes / demand will change rapidly, allocative efficiency considerations could point in favour of using a dynamic, rather than static, benchmark.

Table 10: The relative benefits of static and dynamic approaches

Benchmark	Is likely to be beneficial when...
Static	• You want to attach most weight to technical efficiency. • The pace of technological change is slow to moderate. • Consumer preferences do not change rapidly.
Dynamic	• You want to attach most weight to allocative efficiency. • The pace of technological change is moderate to fast. • Consumer preferences change rapidly.

Source: Economic Insight

5.5.3.2. Controllability of the outcome

An ODI will only incentivise a company to improve when there is a connection between a company's actions and the measure of performance used to set the PC. Without this connection, the ODI would not be effective, and it would not matter whether the PC is static or dynamic.

In practice, there is a spectrum of controllability, and how controllable an outcome is typically rests on:

- first, the extent to which *external factors* (such as the weather) can affect the measure of performance relative to company behaviour; and
- second, *the time horizon* over which performance is measured. That is, where it takes time for companies to make the operational changes needed to affect an outcome measure, **an outcome may be controllable in the long-term, but not in the short-term.**

> *"An outcome may be controllable in the long-term but not in the short-term."*

Therefore, a dynamic benchmark is more likely to make sense when, in addition to seeking to encourage allocative efficiency over technical efficiency (as discussed above), one can be sure that:

- within price control fluctuations (possibly annual fluctuations, depending on the type of dynamic benchmark used) in the performance measure **are a consequence of company behaviour, not external factors;** and
- companies can **make the operational changes needed to affect the outcomes within price control**.

Where these conditions do not apply, a static benchmark would help avoid the risk of companies trying to meet a dynamic PC that is unduly easy or stretching. This is consistent with the position expressed in the recent UKWIR report on outcomes:

"Dynamic PCs could provide an additional incentive for companies to strive for greater service improvements. However, it is also possible that dynamic targets could introduce additional uncertainty. This may result in companies attempting to increase baseline costs to cover any unexpected expenditure, and in the extreme case, this could lead to over-bidding on costs across the industry as a whole. The balance between these two effects suggests that dynamic PCs appear to be most appropriate for low asset areas that are reasonably controllable and can be measured consistently across companies."[56]

[56] *'Setting Performance Commitments and Incentives to Deliver Best Value for Money', UKWIR (2016), page 18.*

5.5.3.3. Uncertainty

A dynamic benchmark gives rise to uncertainty over what a company's PC will be during the price control period. A potential consequence of this uncertainty is that companies may ***'err on the side of caution'*** from the outset, and make investments to reduce the chance of receiving a penalty, or increase the chance of receiving a reward. But this comes with a cost: by revising investment plans to reflect the incentive, companies will divert resources from elsewhere. This could create both allocative inefficiency (precisely what one is trying to avoid by using a dynamic incentive in the first place); and technical inefficiency.

Alternatively, companies may prefer to adopt a ***'wait and see'*** approach. This would reduce the costs associated with making investments that are not ultimately needed, but could instead result in other inefficiencies associated with meeting new short-term targets. For example:

- first, a project with lower fixed / up-front costs and higher variable / on-going costs may be 'least cost' at the lower performance level, but not at the higher performance level; and/or
- second, additional costs are incurred when changing operational plans unexpectedly (e.g. rolling back the existing plans); and or
- third, to mitigate the above, contractual or other contingencies need to be in place from the outset.

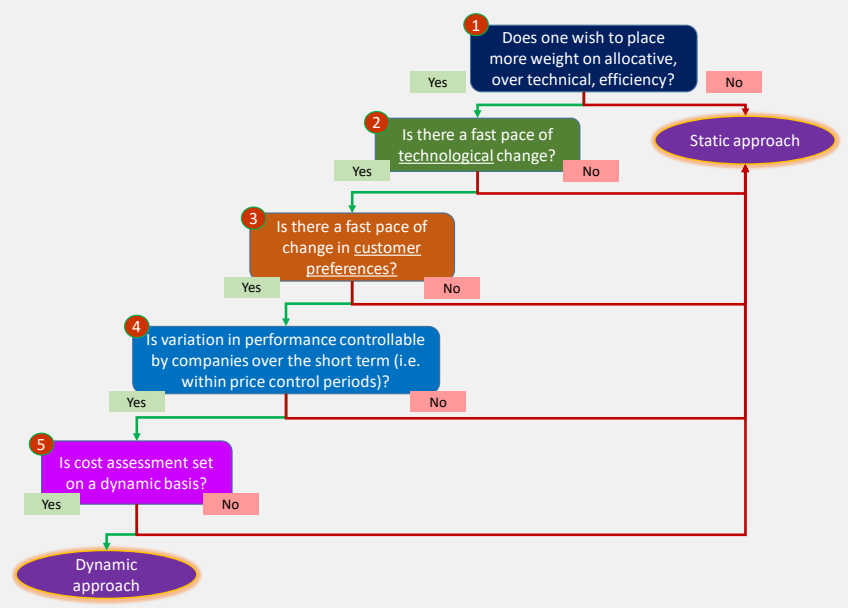
5.5.3.4. Consistency with cost assessment

As we have previously explained, it is essential that the approach to outcomes is consistent with the approach to cost assessment. Logically, therefore, it would not be appropriate to:

- apply a dynamic approach to PCs where a static approach was applied to cost assessment; nor
- apply a static approach to PCs where costs were assessed dynamically.

Drawing the above discussion together, the following figure sets out a decision tree for the choice between static and dynamic benchmarks.

Figure 20: Decision tree for the choice between static and dynamic benchmarks



Source: Economic Insight

5.5.4. Demonstrating dynamic approaches in the water industry

In this section, we consider the relevant empirical evidence relating to the use of different benchmarks in the water industry, focusing on the issue of ***controllability***.

5.5.4.1. Controllability

The table below shows the upper quartile performance levels in each year between 2011 and 2013 that contributed to the three-year average (also shown) used by Ofwat to set the UQ PCs in PR14 for the comparative areas.

Table 11: Changes in UQ performance over time

	2011	2012	2013	Average
Water quality contacts per '000 population				
Upper quartile performance level	1.27	1.32	1.18	1.23
Percentage change		4%	-11%	
Properties subject to sewer flooding per '000 properties				
Upper quartile performance level	0.13	0.17	0.14	0.15
Percentage change		25%	-15%	
Water drinking quality (mean zonal compliance %)				
Upper quartile performance level	99.98	99.97	99.97	99.97
Percentage change		0%	0%	
Pollution incidents per length of sewer km				
Upper quartile performance level	31.89	43.04	40.46	41.56
Percentage change		35%	-6%	
Length of supply interruptions per property				
Upper quartile performance level	10.20	11.70	9.30	10.40
Percentage change		15%	-21%	

Source: Economic Insight analysis of Ofwat 'upper quartile comparative assessments' data.

The table shows that **there are significant year-on-year variations in the level of the performance measures** – north of 10% is not unusual; and up to 35% in the case of pollution incidents.

If a dynamic benchmark was used, such that companies' PCs were set equal to the UQ performance in the previous year, this analysis reveals that there would be significant changes in PCs within a price control period. Indeed, even an attempt to *'smooth'* the PCs by taking an average of the previous two years would seem to result in material changes to PCs.

To the extent that dynamic benchmarks of this type are a realistic option for PR18, it will be important to first understand what causes this volatility; and, in particular, whether it is caused by *external factors outside of company control.* If it is, this would point towards either:

- developing a performance measure that better **distinguishes between changes in company performance and changes in external measure**; or
- **using a static rather than dynamic benchmark** (assuming that average performance over a five-year period is less volatile than it is year-to-year).

5.5.5. Implications for PR19

The discussion and evidence set out in this section suggests the following implications for the PR19 approach.

- **First, choosing between a static benchmark and a dynamic benchmark, involves deciding the appropriate balance between technical and allocative efficiencies.** This balance might vary between outcomes. Dynamic benchmarks are more appropriate for outcomes subject to rapid (within price control) change in technology or consumer preferences. In general, however, water would not typically be characterised as an industry with fast changing customer tastes, nor technological change. Accordingly, this generally indicates that customers will be best protected by the strong technical efficiency incentives of static approaches to benchmarking.
- **Second, dynamic benchmarks are likely to be more appropriate to outcome measures that are 'highly controllable' by companies within a price control period** i.e. in the sense that (a) they are not materially driven by changes in external factors; and (b) companies can make rapid operational changes to meet new PCs. The extent of year-on-year volatility in the comparative performance measures suggests that there is a real need to better understand the extent of (a) and (b) before implementing a dynamic benchmark approach.
- **Third, a dynamic benchmark approach could create uncertainty.** Ofwat should develop an understanding of what companies would need to do to meet 'moving targets' and, relatedly, whether customers would be willing to pay for these activities.
- **Fourth, given our views that a consistent approach is required across cost assessment and outcomes, this rationale should carry across to the choice between static and dynamic benchmarks also.** That is to say, should a static approach continue to apply in relation to cost assessment, this would further imply that a static approach should apply to outcomes – and vice versa.

5.6. The pros and cons of using wider benchmarks

5.6.1. The in principle pros and cons of using wider benchmarks

The term 'wider benchmarks' refers to the potential to set performance targets (for a subset of common PCs) based on comparators *outside of the water industry in England and Wales.* Such wider benchmarks could include both: (i) water companies *in other countries* – i.e. international comparators; and /or (ii) companies in *other industries.*

The main potential benefit of using wider benchmarks is that it might increase the information available to companies and the regulator on what performance levels *could* be achievable.

Indeed, wider benchmarks are necessary where within-sector comparisons are not possible. For example, the ORR has used international comparators when benchmarking Network Rail's performance, *because there is only one such company in the UK;* and, similarly, the CAA has compared Heathrow airport to other large international hubs.

However, the main challenge associated with developing such benchmarks relates to the extent to which they provide a good basis for comparison, or whether the services and/or operating environments in question are simply *too different* to be of practical use. This comparability issue likely explains why **those regulators that are obliged to use wider benchmarks, tend to look to similar industries in other countries, rather than different industries.**

5.6.2. The economics issues associated with applying wider benchmarks

Before using wider benchmarks, it is important to assess how comparable they are against two high-level criteria:

- first, whether the conditions of supply (cost) and demand (consumer preferences) are *similar* between the water sector in England and Wales and the candidate comparator; and
- second, *whether the sector uses similar outcome measures.*

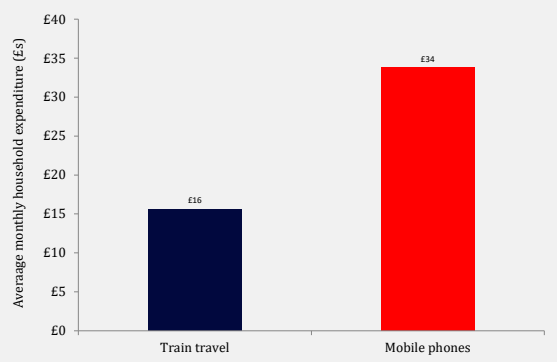
To illustrate the importance of the first point – imagine one was to compare outcomes between industries as diverse as train travel and mobile phone network operators. In principle, perhaps one could identify an outcome measure that *appears* to be broadly similar: say, network reliability (e.g. measured in terms of the number of 'outages' per customer per year).

Critically, however, fundamental differences in both the economically efficient cost of operating those networks, and the value that customers attach to network reliability, could mean that the economically efficient outcome is radically different between the two. In this illustrative example, this could be because:

» **Customers' value** of mobile phone network reliability is driven by the benefits they get from being able to communicate, download data, use apps, and so on. Whereas, their value of train network reliability will be a function of their use of trains and the value they attach to their time.
» **In relation to costs,** mobile phone network reliability is a function of available spectrum bandwidth; and the quality and configuration of masts. In contrast, train network reliability will be a function of expenditure relating to the quality of rail track, points, signals, general maintenance, and so on.

To illustrate just how large differences in 'value' can be across industries, the following figure shows household monthly expenditure in relation to train travel and mobile phones. This reveals that expenditure on phones is double that for train travel.

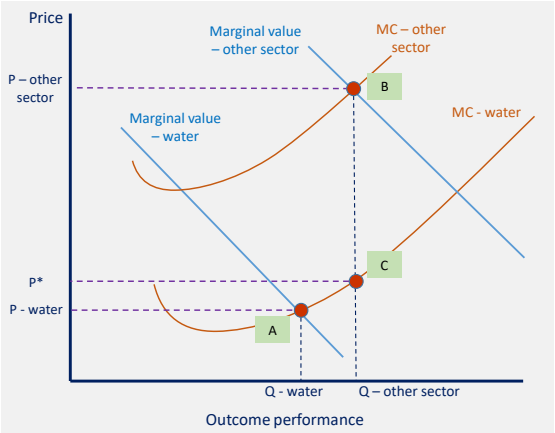
Figure 21: UK average household monthly expenditure



Source: ONS Family Spending survey 2014

Importantly, if one were to apply benchmarks in other sectors where the economically efficient outcome varied materially from water, there would be customer detriment. The following figure illustrates why this would be the case.

Figure 22: illustrating the potential welfare loss from inappropriate comparators



Source: Economic Insight

In the above example the economically efficient outcome performance for water is shown by *'Q – water'*. This is below that economically efficient outcome for another comparator sector *'Q – other sector'*. Accordingly, if one were to set an outcomes benchmark for water based on the comparator, customer detriment would arise because:

- In order to deliver that outcome, water companies would need increased funding (moving from point *A* to point *C* in the figure). However, **were Ofwat to allow this, it would be funding a level of outcome delivery (paid for by customers) that was in excess of customers' valuation** – and so there would be customer harm.
- Equally the outcomes benchmark might be set at the level of the comparator, but with no increase in funding by Ofwat. In that case, **the outcome would not be deliverable, and so there would be no incentive on water companies** (also resulting in customer harm). In addition, over time, companies would experience a funding shortfall.
- Also, **it might be that the level of outcome performance implied by the benchmark is not deliverable at all** (i.e. no matter how much funding is increased by). In this case, equally there would be no incentive power on companies – and thus, customer harm would arise.

The above does not imply that wider benchmarks would never be appropriate. Rather, the purpose here is to illustrate the importance of ensuring that, where such wider comparators are used, that the supply and demand conditions are *sufficiently similar* to water for the analysis to be meaningful (and the resultant incentives to be appropriate).

A further important point of clarification is that **the suitability of wider benchmarks may differ across the water value chain.** That is, there may be particular elements (most likely those relating to retail) where

comparators with 'more similar' demand and supply conditions can be found.

5.6.3. Implications for PR19

It is beyond the scope of this report to rule in (or rule out) other specific comparators. However, the preceding discussion has a number of implications for any approach to PR19. These are as follows:

- The relatively large number of companies in the water industry in England and Wales means that extensive within-sector comparisons are possible and, arguably, this could imply that Ofwat and companies should devote proportionately less time and effort to developing wider benchmarks (relative to cases where such 'within sector' comparisons are not possible).
- Where wider benchmarks are used, water sectors in other countries may be most appropriate in the first instance. However, care would still need to be taken to ensure that they are *comparable* before use. As explained, this would involve an analysis of the relevant demand and supply side conditions.
- It is important, however, to acknowledge that the above is a *'matter of degree',* and that the criteria we have identified might imply differing approaches across certain outcome areas – or, indeed, for different parts of the value chain. For example, **wider approaches might be more applicable in retail** than in wholesale.
- In cases where Ofwat does consider the application of wider (and in particular, non-water) benchmarks, it will be important to undertake a robust comparison of the supply and demand conditions. On the supply side, this should include:
  - a qualitative analysis of the *processes involved in supplying the good or service* in question, and an assessment of 'how similar' those processes are to those involved in the relevant part(s) of the water value chain; and where possible
  - a comparison of the prevailing costs associated with supplying the good or service in question, and how similar those are to the relevant part(s) of the water value chain (this would include analysis of cost structure, such as capital intensity, the extent of sunk and fixed costs, and asset lives).
- On the demand side, where data allows, qualitative and quantitative analysis of how the 'value' to customers of the comparator good or service compares to the relevant part of the water value chain is also important.
- Following from the above, in general it would seem that the most appropriate role for wider benchmarks might not be to set the PCs per se; but rather, to encourage consideration of different ways of meeting environmental and other challenges in the longer-term. Or indeed, to utilise such comparators primarily for the purpose of monitoring and revealing information, rather than setting incentives. Again, however, this too might differ across the value chain.

6. Conclusions and recommendations

In this final section of our report, we forward our conclusions and recommendations. The primary purpose of this is to help provide clear, practical solutions that will assist Ofwat and other stakeholders in the development of a more detailed methodology for outcomes at PR19.

Our main recommendations are as follows:

(i) Recognising the inherent uncertainty in establishing whether the efficient level of outcome performance varies across companies, **there should be a tolerance range for PCs subject to comparative benchmarking.** In turn, Ofwat should nominate common and comparative PCs **in advance of plan submission.**

(ii) To help ensure consistency of funding and outcomes, **there should be a consistent approach to outcomes and cost assessment, with both occurring concurrently.**

6.1. The balance between bespoke and common PCs – and the role of comparative assessment

The choice between bespoke and common PCs (and the application of comparative assessment to a subset of common PCs) should centre on **establishing where the economically efficient level of provision varies, or is 'sufficiently similar', across companies.**

In practice, there will always be a degree of uncertainty regarding the similarity of the efficient level of outcome performance. We therefore advocate:

- **That the outcomes framework for PR19 should be explicitly designed to accommodate this uncertainty** in an appropriate way.
- That the process for evolving the framework **should explicitly seek to mitigate the uncertainty where possible** – thereby increasing the likelihood that appropriate choices are made between bespoke, common and comparative approaches.

Both a transparent process, and high quality evidence, will be essential to make the above possible.

In relation to **mitigating uncertainty,** our recommendations are as follows:

- In the longer-term, a process whereby **some form of industry-wide method** (or set of methods) is deployed to assess customer value would seem to be the only credible way to ascertain whether, and to what extent, there are genuine differences in customer value across companies.
- In the nearer term, and **for PR19, an alternative process to the above is likely to be needed – drawing on the existing (e.g. PR14) WTP evidence base** and other qualitative information.
- To help better understand what might drive differences in the costs of providing certain levels of outcome performance across firms (and to ensure consistency of funding and outcomes) **there needs to be a consistent approach to outcomes and cost assessment.** We expand on this in the next sub-section.

In relation to **accommodating uncertainty** in an appropriate way, our recommendations are:

- In relation to the subset of common PCs to which comparative assessment is applied, **there should be a 'tolerance range' around any benchmarks Ofwat identifies.** This is for two main reasons: (i) to mitigate the risk of consumer harm arising from Ofwat making interventions that, unintentionally, result in setting PCs that are either above, or below, the economically efficient level (which one would never expect to be *exactly the same* across firms); and (ii) because if Ofwat did not intend to allow any variation in company PCs in these cases, there would be no need for companies to propose their own PCs in the first instance (something most stakeholders consider to be valuable).
- **Companies should continue to propose their own PCs,** even in relation to the subset of common PCs subject to comparative assessment. This is both to encourage 'company owned plans', but also because it is consistent with the broader process we have recommended.
- **Ofwat should pre-designate which subset of common PCs will be subject to comparative assessment** prior to plan submissions (by definition, Ofwat will need to identify which are common).
- We also recommend that Ofwat **publish details of its comparative outcome benchmarking,** and the implied PCs, in advance of plan submission (importantly, this is consistent with its approach to cost

assessment). This is because, if Ofwat did not do this, the logical implication is that: (i) the variation in company proposed PCs would be greater at the time of Plan submission; which (ii) would only be valuable to Ofwat if it wanted to set a *wider* tolerance range for those PCs.[57] In addition, by publishing such information in advance of plans, CCGs would be able to play a more meaningful role to challenge companies in relation to the relevant PCs.

- Following from the above, the **responsibility should sit with companies to provide high quality evidence and analysis to support their proposals.** This would include, in relation to any common PCs subject to comparative assessment, **evidence to support any proposed deviation from Ofwat's benchmark** (within the tolerance range).
- Accordingly, **it should be for Ofwat to set the required evidential hurdle**. Therefore, the extent of **any variance in comparative PCs would be at Ofwat's discretion.** Ofwat would apply its evidential hurdle as part of the RBR process.

6.2. How outcomes interact with cost assessment

The approach to outcomes (specifically the setting of benchmarks for a subset of common PCs using comparative assessment) **must be consistent with the approach to cost assessment; and should occur concurrently.** This is to ensure that the level of funding allowed for within the PR19 determinations is consistent with any outcomes set.

In order to achieve this, we recommend the following:

- At present, an approach based on **'ex-post' modelling adjustments that 'adjust' allowed costs to reflect the outcome performance would seem to be preferred.** To achieve that, further analysis to examine how improvements in certain outcomes impact incremental cost should be prioritised. This should include both top-down statistical analysis; and more granular bottom-up cost modelling.
- The process for applying ex-post modelling adjustments should ideally be Ofwat owned, and **administered in a consistent way across all companies** – with details included in any methodology for PR19.

6.3. The choice of benchmark

Ofwat should retain an upper quartile benchmark for the subset of common PCs to which it applies comparative assessment. However, the choice between this, and upper quintile, appears finely balanced; and so this is a matter of judgement. Alternatives including 'pure frontier' or 'average best of three' appear too demanding. Furthermore, again the choice must be consistent with the approach to cost assessment.

In weighing up the choice between upper quartile and upper quintile, Ofwat could analyse 'how achievable' upper quartile proved to be over PR14. This could include, for example, undertaking an evaluation of the totality of the efficiency challenge at PR19 in percentage terms (i.e. using upper quartile) and compare it to:

[57] *The alternative, of companies proposing PCs, but Ofwat not ultimately allowing any variation in those PCs (where comparative benchmarks are set) would clearly be at odds with not publishing information in advance of plans.*

- the % challenges that were applied at previous price controls in the sector; and
- the % challenges that have been applied across other regulated industries.

The above would provide a robust view on 'how challenging' the upper quartile (or other benchmarks) are compared to 'adjusted residuals' approaches, and so on.

6.4. Whether benchmarks should be set 'individually' or on an 'aggregate' basis

There is insufficient evidence to reach a strong view as to whether individually set benchmarks, or an aggregate benchmark, across comparative outcomes measures is most appropriate. The decision turns on establishing the extent to which there are 'trade-offs' across the individual areas. However, **what evidence there is points to giving serious consideration to an aggregate approach.** We note that, were Ofwat to take a more aggregate approach, it might be sensible to set a 'more demanding' challenge relative to an individual approach, reflecting the benefit of the flexibility this gives companies.

In addition to understanding variation in customer value across companies (described earlier), we would recommend taking forward work to understand:

- whether there are inherent trade-offs in outcome performance at a granular / individual outcome level; and
- whether the costs of delivering outcomes are 'more similar' in aggregate than they are 'individually' across companies.

6.5. Applying static or dynamic benchmarks

Based on the available evidence, the technical efficiency loss from using dynamic benchmarks seems likely to outweigh any allocative efficiency gain in the water sector. Pending any better evidence, **Ofwat should therefore retain a static approach to setting outcomes benchmarks – as this appears to be in the best interests of customers.** The choice between static and dynamic approaches should also reflect the principle of consistency with cost assessment. That is to say, as a matter of principle, Ofwat should not deviate from a static approach to outcomes for so long as it retains a static approach to cost assessment – and vice versa.

As present, **the data shows considerable year-to-year volatility in the comparative benchmark measures.** This, amongst other things, is most consistent with customers benefitting from a 'static' approach. In order to even consider dynamic approaches, further work is needed to understand: (i) the extent to which this volatility is driven by external factors, outside of company control; and (ii) the extent to which companies in practice are able to make year-to-year operational changes that would allow them to meet dynamically set PCs (without there being undue trade-offs).

6.6. The scope for using wider (e.g. international, or other industry) benchmarks

First principles suggest that **in many cases, within industry benchmarking is likely to be most appropriate** – with some scope for international benchmarking also within the water industry. We note, however, that the appropriateness of wider benchmarks outside of the water sector fundamentally rests on establishing 'how similar' the

relevant demand and supply side conditions are across any comparators. Accordingly, the suitability of such wider comparisons might vary by outcome type, and across the value chain. Intuitively, **they might be more appropriate for retail,** for example.

Generally, across most outcome areas, this issue should not be a priority. However, to the extent that wider benchmarks are a consideration in relation to some outcomes, detailed, robust, evidence should be developed to inform how 'comparable' they are.

It might also be that wider comparators are most appropriate for 'informational' purposes, rather than for setting incentives. In which case, the above issues may be less of a concern.

6.7. Concluding remarks

The recommendations set out here are consistent with the overarching framework we identified regarding the role and purpose of the outcomes framework for water. Further, they are designed to work as a coherent whole, which recognises the close interdependencies that exist across the multiple policy issues the industry is seeking to address.

We recognise that not all stakeholders will agree with all of our specific recommendations. Further, in a number of areas we make it clear that additional evidence and work would be required in order to reach more definitive views. Nonetheless, we hope that, by providing clarity around the key issues (and the factors that should determine one's assessment of them) our report will help draw out the reasons for any areas of agreement and disagreement. In turn, the hope is that this will help support policy development in this vital area, allowing the industry to build on the widely recognised successes of the PR14 outcomes framework.

7. Appendix A: re-cap of the PR14 approach to outcomes

Ofwat's new approach to outcomes at PR14 encouraged companies to focus on delivering against customer priorities

At PR14 Ofwat introduced a new outcomes framework, where emphasis was placed on creating greater incentives for companies to deliver the outcomes their customers wanted. A key element of this was the introduction of Outcome Delivery Incentives (ODIs). In its Final Methodology document for PR14[58], Ofwat described its intended approach to ODIs as having the following key features:

- **A flexible approach.** In particular, Ofwat stated that it was for companies to propose their own ODIs, with associated performance commitments (PCs) – and that these could therefore vary across companies, to reflect differences in customer priorities.
- That the **incentives should reflect the value customers place on outcomes**, supported by robust willingness to pay (WTP) evidence.
- That companies should propose **both penalties and rewards** within the flexible framework. Penalties should apply where companies fail to deliver on their PCs, and rewards should apply where there is outperformance.

Whilst Ofwat placed emphasis on 'flexibility', within its final methodology the regulator also stated that there were circumstances where having 'consistent' outcomes across companies made sense. Specifically, Ofwat stated that this might be the case where: (i) an incentive mechanism worked on the basis of comparison across companies and / or (ii) the outcome in question is known to be of significance to all customers, or the environment. Accordingly, Ofwat mandated two 'consistent' outcomes across the companies in relation to:

- leakage; and
- SIM.

Reflecting Ofwat's 'flexible' approach, its original intention was to review outcomes within its Risk Based Review (RBR) process

In keeping with its stated objective of giving the companies flexibility, Ofwat's intention was to primarily play the role of 'reviewer' in relation to outcomes. Specifically, company proposed-outcomes were to be incorporated within Ofwat's broader RBR process, which ran from January to March 2014. It was intended that the RBR process should provide scrutiny across 3 areas relevant to outcomes: (i) customer engagement; (ii) PCs; and (iii) the ODIs. In all cases, the RBR included reviewing company plans against criteria relating to: process, method and the results themselves.

The key point here, was that, with the exception of the two 'consistent' ODIs, Ofwat wanted companies to retain ownership of outcome incentives. This meant it was not for Ofwat to set

[58] *'Setting price controls for 2015-20 – final methodology and expectations for companies' business plans.' Ofwat (July 2013).*

ODIs (either ex ante or ex post) – but rather, the regulator was simply there to scrutinise and challenge the outcomes.

Ofwat had a number of serious concerns with company ODIs at the time of Plan submission

When companies submitted their plans, Ofwat found that, whilst the approach to customer engagement was broadly positive, there were serious concerns relating to the PCs and ODIs. Key concerns Ofwat identified were:

- overreliance on non-financial incentives;
- absence of financial rewards;
- failure to link customer WTP with the scale of incentives;
- inappropriate scaling of incentives;
- inappropriate use of neutral zones (or dead-bands); and
- failure to explain how ODIs were calibrated with other performance incentives.

Building on the above, Ofwat was particularly concerned that, whilst the greater emphasis on 'penalties' over 'rewards' appeared to reflect customer preferences, this might be because customers did not properly understand the implications of this for the WACC (i.e. it could mean higher bills).

The issues around the calibration of ODIs were magnified by there being very large differences in WTP across companies, for relatively comparable metrics. For example, the differences in WTP to avoid sewer flooding ranged from £25,540 to £434,319 per property. Put simply, Ofwat was concerned that customers of different companies might pay substantially different amounts for similar outcomes.

Taken in the round, the above concerns led Ofwat to conclude that company plans did not adequately balance risk and reward for customers. Consequently, in April 2014 Ofwat issued risk and reward guidance, which included specific steps in relation to ODIs, where Ofwat stated: *"companies should resubmit ODIs with meaningful rewards and penalties as part of this package."*[59] Also in April 2014, Ofwat held a company workshop specifically on ODIs, where it provided further details of its concerns.

Ultimately, Ofwat followed a very different process from the one it originally intended for outcomes, with ODIs being fully de-scoped from the RBR. The evolution of Ofwat's process is illustrated in the figure below.

Figure 23: Evolution of PR14 outcomes process

[59] *'Setting price controls for 2015-20 – risk and reward guidance.' Ofwat (January 2014); page 4.*

Ofwat made substantial interventions in company proposed ODIs

Having removed ODIs from the RBR process, Ofwat made direct interventions to company plans at both the Draft and Final Determination stage of PR14 in relation to outcomes. In broad terms, there were four key types of intervention:

- **Upper quartile.** Where companies proposed similar/ comparable outcomes measures, Ofwat undertook *comparative analysis* in order to re-calibrate proposed ODIs so that: (i) companies faced penalties in the event of non-delivery; and (ii) could only benefit from rewards where their performance exceeded upper quartile levels (by 2017/18).
- **Company specific calibration.** Where cross-company comparisons were not possible, Ofwat performed company-specific analysis that included: (i) checking compliance with Ofwat's methodology; (ii) assessing 'value for money'; and (iii) protecting customers against under-delivery.
- **Cap and collar.** Ofwat imposed an aggregate 'cap and collar' on the incentives associated with ODIs. These were equivalent to +/-2% of RORE. The cap applies at the end of five years, allowing netting off year-to-year; and is applied separately for water and wastewater. In some instances, specific ODIs were excluded from the cap / collar.
- **Deadbands.** In some cases, Ofwat imposed 'deadbands' on ODIs, which refer to limited variations in performance around targets where no incentives apply. This was to reflect the fact that there could be minor variations in outcomes that are unrelated to company performance and / or are outside of company control. Deadbands apply for the first two years of the control only, to provide some transitional protection against penalties for falling below upper quartile performance during this time.

In its interventions, Ofwat applied a 'comparative' approach across the following areas:

- duration of supply interruptions;
- number of contacts from customers regarding quality of water;
- compliance with DWI water quality standards;
- number of sewerage pollution incidents;
- number of properties impacted by internal sewer flooding; and
- leakage.

Ofwat primarily identified these areas based on 'coverage'. That is to say, Ofwat considered it more appropriate to apply a comparative approach for outcomes where all, or most, companies had proposed a related ODI.

In order to then implement its calibration of comparative ODIs based on upper quartile performance, Ofwat further:

- normalised relevant data in order to control for differences in company size;
- where companies proposed metrics other than those Ofwat used for making comparison, Ofwat 're-stated' the company's proposed metric; and
- Ofwat had to collect and adjust historical performance data in order to reach a view on what the 'upper quartile' level of performance was for each area.

In relation to the above, we note that at various stages in the process Ofwat asserted that companies were funded for upper quartile service performance – but that companies contested this point. For example, the following extract is taken from the submission of South East Water's CCG to the draft determinations:

"SEW initially advised the CCG that companies are not funded by Ofwat to achieve upper quartile performance because the totex modelling is based on average cost. When the CCG Chair raised this at the Ofwat workshop on 11 September 2014 and when SEW raised this with the Ofwat ODI team on 15 September 2014, Ofwat advised that efficiency reflects a combination of both what it costs to deliver and how much is delivered (i.e. the quantity or quality). So, in Ofwat's view, the principle that companies are being funded to upper quartile levels of efficiency covers both costs and performance.

SEW contests this and the CCG can understand the Company's concern here. It is clear that the funding is for average performance (the models use historical average unit cost and apply it to activity forecasts for 2015-20, activity required to maintain levels of service) and the efficiency challenge applied is based on whether a company's costs represent upper quartile efficiency."[60]

Similarly, in its representations Thames Water stated:

"Ofwat has assumed that customers have paid for upper quartile performance. However, most customers will neither have paid for upper quartile performance in past bills, nor would they pay for this during 2015-20 because Ofwat's cost models determine the upper quartile efficiency for average performance, not for upper quartile performance."[61]

In addition to the above, the details of 'how' Ofwat identified the relevant upper quartile for each outcomes area was also a source of debate amongst the companies.

[60] *'South East Water Customer Challenge Group Submission to Ofwat on the Company's Draft Determination for 2015-2020.' (2nd October 2014).*

[61] *'Draft Price Control Determination: Thames Water Response.' Page 11.*

8. Appendix B: evidence on cost quality trade-offs

8.1. Academic literature review – theory and empirical evidence

The presence of cost / quality trade-offs is well established in the economics literature. In the table below we summarise the key evidence we have reviewed.

Table 12: Literature review of cost / quality trade-offs

Title	Author(s)	Journal	Key findings of relevance
Empirical studies			
The Trade-Off between Costs and Outcomes: The Case of Acute Myocardial infraction	Jonas Schreyögg and Tom Stargadt	Health Services Research	Costs were negatively associated with mortality and readmissions. Every U.S.$100 less spent is associated with a 0.63 percent increase in the hazard of dying and a 1.24 percent increase in the hazard to be readmitted conditional on not dying. Suggesting that there is a trade-off between costs and outcomes.
Interrelationships Between cost Overrun and Quality of Provincial Public Section Development Projects Analysing Dynamics Involved in Provincial Projects of Punjab	Ali Adnan Joiya, Rana Saifullah Hassan, Ali Zeshan Joiya and Dr Anwar Hassnain.	Journal of Culture, Society and Development	This study revealed that there is highly positive and significant relationship between costs and quality of provincial public section development projects.
Exploring the relationship between costs and quality in Danish hospital department	Anne Hvenegaard, Jacob Nielsen Arendt, Andrew Street, and Dorte Gyrd-Hansen.	University of Southern Denmark – Health Economics	The association between cost and quality differs depending on how quality is measured. Lower costs tend to associate with higher mortality, implying a cost-quality trade-off. In contrast there is no clear association between costs and wound complications among vascular departments.
Relationship between quality and cost: pure and simple?	Steven Fleming.	US National Library of Medicine Nation Institutes of Health	Results show several quality measures to be statistically significant determinants of cost. With each measure. The cost-quality relationship is nonlinear but not monotonically increasing throughout the entire range of quality.
The trade-off between hospital cost and quality care. An exploratory empirical analysis	Morey RC, Fine DJ, Loree SW, Retzlaff-Roberts DL, Tsubakitani S.	Med Care	The marginal additional cost per each death deferred in 1983 was estimated to be

			approximately $29,000 (in 1990 dollars) for the average efficient hospital. Also, over a feasible range, a 1% increase in the level of quality of care delivered was estimated to increase hospital cost by an average of 1.34%. This estimated elasticity of quality on cost also increased with the number of beds in the hospital.
Product Customisation and Customer Service Costs: An Empirical Analysis	Anuj Kumar and Rahul Telang	Manufacturing and Service Operations Management	Shows that customers with a customized plan call 21% less frequently than customers with a standard plan. They also show no difference in the claims adjudication cost between a standard and a customised plan exists. Overall, the results suggest customised plans may be operationally cheaper to serve than standard plans and thus provides a link between a growing business concern (customer support via call centres) and a prevalent business strategy (product customisation).
Theory studies			
The role and measurement of quality in competition analysis	The OECD	Directorate for Financial and Enterprise Affairs Competition Committee	In response to the entry of the low cost carriers, incumbent airlines adopted a number of different strategies. First, some airlines engaged in tariff mimicry, that is, they reduced their own fares in an attempt to match the prices charged by the new entrants. The "Southwest effect," named for the first low cost airline that emerged in the United States in the 1970s, demonstrates the severe impact of low cost carrier entry on the fares charged by established operators. Second, some full service airlines focused upon improving the quality, and in particular the punctuality, of their own offerings. This is similar to the effects seen in the retail groceries sector, where entry by an aggressive price competitor tends to result in an improvement in stock levels in neighbouring rival stores. In France, for example, some retail grocery chains have chosen to reposition themselves as higher service outlets in

			response to market entry by deep discounters such as Lidl and Aldi.
Competitive Advantage: Creating and Sustaining Superior Performance	Michael Porter	Free Press	Discusses the choices firms face in terms of pursuing cost leadership or differentiation and the trade-offs associated with these options.
The role and measurement of quality in competition analysis	The OECD	Directorate for Financial and Enterprise Affairs Competition Committee	It is not the case, for example, that an increase in market concentration necessarily leads to a decrease in product quality or choice. Moreover, market entry by low cost airlines is expressly premised, on lower (or more basic) levels of service in exchange for lower ticket prices. Yet there is some evidence of cost carriers' service being generally more successful at delivery of one specific quality attribute, punctuality, which air passengers may view as indispensable – depending on the particular market at issues, therefore, the empirical research indicates that changes in competition levels can have either positive or negative effects on quality.
The welfare effects of airline deregulation in the united states	Donald William Koran	Federal Trade Commission, Washington, D.C	After deregulation, airline profits did not fall, but they lowered fares – Airlines did not really lower fares, but instead offered different services that were cheaper to produce. With deregulation, fares fell by more than enough to offset the deterioration in service quality. Producer's surplus remained substantially unchanged, because the fall

			in fares was roughly matched by a decline in average costs associated with lower service quality.

8.2. Our cross-industry analysis of cost / quality trade-offs

As outlined in the main body of our report, one expects there to be clear trade-offs between cost and quality in competitive markets. Accordingly, we collected a range of data to enable us to test this across a number of industries that would broadly be accepted as being 'competitive'.

In order to examine the extent there is trade-offs between cost and quality we tested these relationships for both the car manufacturing and airline sectors. Two industries that are broadly considered to be competitive. To do this, we applied the following approach:

- firstly, we used statutory accounting data and annual reports in order to identify a range of measures of cost for each firm (i.e. operating costs) within the industry in question;
- following from the above, we then developed some broadly comparable measure of unit cost for each firm (i.e. cost per passenger km); and lastly
- we identified a third party measure of 'quality' for the firms in question.

In the following we provide further details of our method, data and findings.

8.2.1. Car manufacturing

Our approach to testing the relationship between cost and quality in the car manufacturing sector is set out in the following:

» **Quality measure.** We obtained our measure of quality from the JD Power survey,[62] which provides information on car manufacturing performance across four main areas: (i) customer service; (ii) dependability; (iii) quality; and (iv) performance and design. For the purpose of this analysis, we used rating score for the **overall performance and design.**

» **Unit cost.** Here, as a proxy for cost, we used the price per car as listed on Parkers,[63] a well-recognised car comparison website. Where possible, for consistency we collected pricing information for saloon cars, that were also included in quality measure. Where a price range was give, we simply took the midpoint.[64]

» **Car manufactures.** We have included 28 car manufactures in our analysis. To be included, two criteria needed to be satisfied: (i) rated in the JD Power survey; and (ii) have pricing information available on Parkers.

Our findings show that overall, lower priced (cost) cars are typically low quality, and higher cost cars, are of higher quality. Results are illustrated in the following figure.

[62] *www.jdpower.com/ratings/industry/automotive*

[63] *www.Parkers.co.uk*

[64] *We recognise that price isn't a true reflection of the 'cost', as it includes margins, we consider it to be suitable for our purposes considering the time it would have taken to collect unit cost information.*

Figure 24. Relationship between unit price and service level performance across car manufacturers

Source: Parkers and JD Power Survey and price data

8.2.2. Airlines

Our approach to testing the relationship between cost and quality for airlines is set out in the following:

- **Quality measure.** We obtained our measure of quality from SkyTrax ratings[65] (the international air transport rating organisation), which operates a global star rating system for airlines. These include providing, not only overall quality ratings, but also ratings for a range of performance categories, such as (i) the on-board product; (ii) cabin staff; and (iii) airport service, to name a few. For our purposes we have used the overall star rating.
- **Unit cost.** In order to calculate the unit cost measure, we divided operating costs by revenue passenger kilometres (number of passengers carried X by distance flown), for each airline.
- **Airlines.** We have included four of the seven UK airlines rated on SkyTrax, these are: British Airways, Easyjet, Virgin Atlantic and Jet2.com. In our analysis, to be included, three criteria needed to be satisfied: (i) rated on SkyTrax; (iii) a UK based airline; and (iii) have the revenue passenger kilometres measure available.

Our findings show that overall, low cost airlines are typically lower quality, than high cost airlines, who ultimately score higher in terms of quality. Results are illustrated in the following table.

Table 13. relationship between cost per passenger km and SkyTrax quality rating

	Operating cost £m (2015)	Passenger km (m)	Cost per passenger km	SkyTrax rating
British Airways	£10,069	142,016	0.07	4
Virgin Atlantic	£2,775	37,157	0.07	4
Easyjet	£4,000	77,619	0.05	3
Jet2.com	£1,203	73,547	0.02	3

Source: company accounts, SkyTrax and statista.com[66]

[65] *www.airlinequality.com/ratings/a-z-airline-rating/*

[66] *Statista.com was used to obtain the passenger kilometres for Jet2.com as these were not provided in their company accounts.*

9. Appendix C: an econometric approach to controlling for outcome performance

This appendix sets out the results of the econometric analysis we have conducted to investigate the feasiblity of including quality performance measures in the econometric benchmarking models used at PR14.

9.1. Ofwat's PR14 cost models

Our start point was the cost models and data Ofwat used in cost assessment at PR14.

In developing its water and sewerage econometric cost models (where Ofwat worked with CEPA), Ofwat used a range of models and approaches to arrive at its cost allowance (using a 'triangulation' process). For the purpose of this work, we selected a subset of the models used.

- **In relation to water** we used an overall refined water totex model, specifically '**model WM5**' as descirbed the CEPA's report[67] (i.e. totex refined translog COLS) and the random effects equivalent **'model WM6'**.
- **In relation to wastewater** we used a wholesale base capex + opex model, specifically **'model SW10'** (i.e. refined translog COLS), which includes treatement and sludge (base) and the random effects equivalent **'model SW9'**.

Our next step was to replicate the above models, to ensure we began from the same results. The original model results for both water and waste water are as follows:

Table C1: COLS (Model WM5 water)

Source	SS	df	MS	Number of obs =	90
				F(11, 78) =	753.38
Model	102.741117	11	9.34010154	Prob > F =	0.0000
Residual	.967010561	78	.012397571	R-squared =	0.9907
				Adj R-squared =	0.9894
Total	103.708127	89	1.16525986	Root MSE =	.11134

ltotex	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
llength	2.855712	.5069389	5.63	0.000	1.846474	3.86495
ldensity	.7449009	1.979015	0.38	0.708	-3.195015	4.684817
llength2	-.0225912	.0200328	-1.13	0.263	-.0624735	.0172911
ldensity2	1.066746	.2754863	3.87	0.000	.5182945	1.615197
llengthdensity	.5122175	.0856519	5.98	0.000	.3416977	.6827374
trend	-.0067462	.0102506	-0.66	0.512	-.0271537	.0136612
lwage	.7195686	.346201	2.08	0.041	.0303353	1.408802
lpopulation_density	.989234	.4888914	2.02	0.046	.0159261	1.962542
lpercentage_relined	.065015	.0202917	3.20	0.002	.0246174	.1054127
lpercentage_resevoirs	-.0139713	.0070573	-1.98	0.051	-.0280213	.0000787
lpercentage_abstractions	.0201406	.0061647	3.27	0.002	.0078676	.0324136
_cons	-15.25163	4.372493	-3.49	0.001	-23.95659	-6.546663

[67] *'Ofwat cost assessment – advanced econometric models' Cambridge Economic Policy Associations Ltd (March 2014).*

Table C2: COLS (model SW10 wastewater)

Source	SS	df	MS		
Model	16.7106161	8	2.08882701		
Residual	.403383173	61	.006612839		
Total	17.1139993	69	.248028975		

Number of obs	=	70			
F(8, 61)	=	315.87			
Prob > F	=	0.0000			
R-squared	=	0.9764			
Adj R-squared	=	0.9733			
Root MSE	=	.08132			

ltotex	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lload	14.65837	1.830235	8.01	0.000	10.99859	18.31815
ldensity	59.13957	18.07899	3.27	0.002	22.98839	95.29074
lwage	.8464545	.4173248	2.03	0.047	.0119619	1.680947
lproportion_treated	.1271102	.035706	3.56	0.001	.0557117	.1985087
lload2	.1020832	.0346995	2.94	0.005	.0326973	.1714691
ldensity2	-1.210682	1.806811	-0.67	0.505	-4.823624	2.40226
lloaddensity	-3.789837	.3923825	-9.66	0.000	-4.574454	-3.00522
trend	.0200605	.0068606	2.92	0.005	.006342	.0337791
_cons	-224.2043	49.48007	-4.53	0.000	-323.1458	-125.2628

Following the above, we then proceeded to modify the above models to include a range of variables relating to the quality performance levels of the companies. In order to do this, our primary source of data was Ofwat's 'Final Determinations for Upper Quartile Comparative Assessments'[68] which includes the underlying data Ofwat used to calculate the comparative performance commitments set at PR14. In the following we set out our approach to including 'quality' performance measures into these cost models.

9.2. Water models including quality performance measures

Performance measures added into the modified water WM5 model were: (i) water quality contacts; (ii) supply interruptions; and (iii) mean zonal compaliance.

In order to maintain consistency with Ofwat's approach, we took the natural log of the above performance measures prior to including them within the model. The results of these inclusions are set out in the table below.

Table C3: COLS (Model WM5 + water quality contacts, supply interruptions and mean zonal compliance outcome measures)

[68] http://www.ofwat.gov.uk/publications/final-determinationsupper-quartile-comparative-assessments/

Source	SS	df	MS		
Model	103.104973	14	7.36464089	Number of obs	= 90
Residual	.603154943	75	.008042066	F(14, 75)	= 915.76
				Prob > F	= 0.0000
Total	103.708127	89	1.16525986	R-squared	= 0.9942
				Adj R-squared	= 0.9931
				Root MSE	= .08968

ltotex	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
llength	4.763818	.595904	7.99	0.000	3.576716	5.95092
ldensity	-4.504534	1.813263	-2.48	0.015	-8.116739	-.8923285
llength2	-.0706578	.020401	-3.46	0.001	-.1112986	-.030017
ldensity2	.8142763	.2466886	3.30	0.001	.3228475	1.305705
llengthdensity	.9194898	.118211	7.78	0.000	.6840014	1.154978
trend	-.006986	.0088837	-0.79	0.434	-.0246833	.0107114
lwage	.3241573	.3526178	0.92	0.361	-.3782934	1.026608
lpopulation_density	-.8100402	.4792649	-1.69	0.095	-1.764785	.1447045
lpercentage_relined	.0528906	.0166099	3.18	0.002	.0198019	.0859792
lpercentage_resevoirs	-.0004553	.0118763	-0.04	0.970	-.024114	.0232034
lpercentage_abstractions	.0430317	.0079106	5.44	0.000	.027273	.0587905
lwater_quality_contacts	-.2387896	.0686665	-3.48	0.001	-.3755802	-.101999
lsupply_interruptions	.1286084	.0329403	3.90	0.000	.0629879	.1942289
lmean_zonal_compliance	217.5802	111.9358	1.94	0.056	-5.407408	440.5678
_cons	-1030.419	515.5132	-2.00	0.049	-2057.374	-3.464069

The results show that the water quality coefficient has the correct sign and is statistically significant, but the supply interruptions coefficient has the wrong sign and the mean zonal compliance coefficient is insignificant at the 5% level.

We then ran the random effects version of the model (WM6) including the performance measures, below we set out the results.

Table C4: RE (Model WM6 + water quality contacts, supply interruptions and mean zonal compliance outcomes measures)

Random-effects ML regression	Number of obs	=	90
Group variable: id	Number of groups	=	18
Random effects u_i ~ Gaussian	Obs per group: min	=	5
	avg	=	5.0
	max	=	5
	LR chi2(13)	=	131.79
Log likelihood = 121.32617	Prob > chi2	=	0.0000

ltotex	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
llength	3.19394	.8733716	3.66	0.000	1.482164	4.905717
ldensity	-1.959458	3.139653	-0.62	0.533	-8.113066	4.194149
llength2	-.0277322	.0302239	-0.92	0.359	-.0869699	.0315055
ldensity2	.7050539	.4304661	1.64	0.101	-.138644	1.548752
llengthdensity	.612898	.1866587	3.28	0.001	.2470536	.9787424
trend	-.0050394	.005969	-0.84	0.399	-.0167383	.0066596
lwage	.7708259	.2973075	2.59	0.010	.188114	1.353538
lpopulation_density	.0155343	.5519379	0.03	0.978	-1.066244	1.097313
lpercentage_relined	.0533611	.0097187	5.49	0.000	.0343128	.0724095
lpercentage_resevoirs	-.0181013	.0144417	-1.25	0.210	-.0464066	.0102039
lpercentage_abstractions	.0157448	.0083968	1.88	0.061	-.0007126	.0322023
lwater_quality_contacts	-.0381544	.0997819	-0.38	0.702	-.2337234	.1574146
lsupply_interruptions	.1206317	.0496388	2.43	0.015	.0233414	.2179219
lmean_zonal_compliance	0	(omitted)				
_cons	-19.83752	7.29022	-2.72	0.007	-34.12609	-5.548954
/sigma_u	.0792125	.0152204			.054355	.115438
/sigma_e	.0480785	.0040794			.0407126	.0567772
rho	.730783	.0856372			.5425313	.8693986

Likelihood-ratio test of sigma_u=0: chibar2(01)= 52.00 Prob>=chibar2 = 0.000

The results show that the water quality coefficient becomes statistically insignificant, the supply interruptions coefficient continues to have the wrong sign and the mean zonal compliance coefficient is omitted due to collinearity.

One possible reason for the odd results is that the quality performance measures are highly correlated with each other or other variables included in the model. The correlation matrix below shows that some of the outcomes, were highly correlated with various cost variables, such as length, density, and wages, these are illustrated in the following table.

Table C5: Correlation of variables

	ltotex	llength	ldensity	llength2	ldensi~2	llengt~y	trend	lwage	lpopul~y	lperce~d	lperc~rs	lperc~ns	lwater~s	lsuppl~s	lmean_~e
ltotex	1.0000														
llength	0.9732	1.0000													
ldensity	0.0576	-0.0860	1.0000												
llength2	0.9713	0.9989	-0.0832	1.0000											
ldensity2	-0.0404	0.0974	-0.9986	0.0942	1.0000										
llengthden~y	-0.6778	-0.7898	0.6772	-0.7869	-0.6842	1.0000									
trend	-0.0215	0.0025	0.0204	0.0025	-0.0200	0.0102	1.0000								
lwage	0.0247	-0.1174	0.5965	-0.1274	-0.5835	0.4484	-0.2583	1.0000							
lpopulatio~y	0.2229	0.0723	0.6964	0.0588	-0.6876	0.3806	0.0653	0.5983	1.0000						
lpercentag~d	-0.0559	-0.1348	0.2172	-0.1349	-0.2067	0.2337	-0.0623	-0.0848	0.2058	1.0000					
lpercenta~rs	0.3477	0.4861	-0.6034	0.4800	0.5905	-0.7230	-0.0067	-0.6920	-0.3901	-0.1530	1.0000				
lpercenta~ns	0.0420	-0.0097	0.3952	0.0044	-0.4077	0.2503	-0.0394	0.0451	0.1353	0.0614	-0.1337	1.0000			
lwater_qua~s	0.0542	0.1630	-0.7275	0.1605	0.7212	-0.5480	-0.0000	-0.7280	-0.5338	-0.1145	0.8117	-0.0049	1.0000		
lsupply_in~s	0.5177	0.5360	-0.4708	0.5268	0.4833	-0.6787	0.0000	-0.2267	0.0120	-0.0387	0.6020	-0.2723	0.4472	1.0000	
lmean_zona~e	0.0942	0.0035	0.0603	-0.0024	-0.0361	0.0277	0.0000	0.2373	0.2192	0.2075	-0.4450	-0.1932	-0.3420	0.1198	1.0000

9.3. Wastewater models including quality performance measures

For wastewater, we followed the similar process to that of water above. The quality performance measures added into the modified sewerage SW10 model included: (i) pollution incidents; and (ii) internal sewer flooding.

As with water, to maintain consistency with Ofwat's approach, we took the natural log of wastewater outcome variables prior to including them within the model. The results of the performance measure additions are illustrated in the following table.

Table C6: COLS (model SW10 + pollution incidents and internal flooding outcomes)

Source	SS	df	MS		
Model	16.8810606	10	1.68810606		
Residual	.232938698	59	.003948114		
Total	17.1139993	69	.248028975		

Number of obs	=	70			
F(10, 59)	=	427.57			
Prob > F	=	0.0000			
R-squared	=	0.9864			
Adj R-squared	=	0.9841			
Root MSE	=	.06283			

ltotex	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lload	4.934578	2.192809	2.25	0.028	.5467778	9.322378
ldensity	5.115328	16.26916	0.31	0.754	-27.43919	37.66984
lwage	1.70853	.3521159	4.85	0.000	1.003947	2.413112
lproportion_treated	.5081798	.0653056	7.78	0.000	.3775037	.6388559
lload2	-.0789947	.0464788	-1.70	0.094	-.1719985	.0140092
ldensity2	.0967395	1.563261	0.06	0.951	-3.031338	3.224817
lloaddensity	-.418283	.6833021	-0.61	0.543	-1.785567	.9490013
trend	.0282582	.0055436	5.10	0.000	.0171655	.0393508
lpollution_incidents_ave	-.0500828	.0426136	-1.18	0.245	-.1353524	.0351868
linternal_sewer_flooding_ave	.4093863	.0623388	6.57	0.000	.2846467	.5341259
_cons	-47.66508	46.77594	-1.02	0.312	-141.2635	45.93336

The results show that the pollution incidents coefficient is statistically insignificant and the internal sewer flooding coefficient has the wrong sign. We found the same results from the random effects version of the model, as shown in the table overleaf.

We also note that the inclusion of the performance measures resulted in large swings in the coefficients on other variables in the model. For example, the coefficient on wages more than doubled, from 0.72 to 1.71.

Table C7: RE (model SW9 + pollution incidents and interflooding outcomes)

```
Fitting constant-only model:
Iteration 0:   log likelihood = -1969.3204
Iteration 1:   log likelihood = -1068.1758
Iteration 2:   log likelihood = -577.07447
Iteration 3:   log likelihood = -311.47224
Iteration 4:   log likelihood = -170.95086
Iteration 5:   log likelihood = -99.866696
Iteration 6:   log likelihood = -66.904116
Iteration 7:   log likelihood = -54.026376
Iteration 8:   log likelihood = -50.484259
Iteration 9:   log likelihood = -50.019217
Iteration 10:  log likelihood = -50.006256
Iteration 11:  log likelihood = -50.006242

Fitting full model:
Iteration 0:   log likelihood =  99.916092
Iteration 1:   log likelihood =  100.31769
Iteration 2:   log likelihood =     100.58
Iteration 3:   log likelihood =  100.59286
Iteration 4:   log likelihood =  100.59287

Random-effects ML regression                    Number of obs      =        70
Group variable: id                              Number of groups   =        10

Random effects u_i ~ Gaussian                   Obs per group: min =         7
                                                               avg =       7.0
                                                               max =         7

                                                LR chi2(10)        =    301.20
Log likelihood  =  100.59287                    Prob > chi2        =    0.0000
```

ltotex	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lload	4.829839	2.242653	2.15	0.031	.4343187	9.225359
ldensity	3.261576	16.73545	0.19	0.845	-29.5393	36.06245
lwage	1.562811	.3737288	4.18	0.000	.8303161	2.295306
lproportion_treated	.4839475	.0761721	6.35	0.000	.3346529	.6332421
trend	.026765	.0053541	5.00	0.000	.0162711	.0372588
lload2	-.0706664	.0459043	-1.54	0.124	-.1606372	.0193044
ldensity2	.3519555	1.567488	0.22	0.822	-2.720264	3.424175
lloaddensity	-.4458674	.6833654	-0.65	0.514	-1.785239	.8935042
lpollution_incidents_ave	-.0397038	.0487247	-0.81	0.415	-.1352024	.0557948
linternal_sewer_flooding_ave	.3920581	.0680945	5.76	0.000	.2585954	.5255209
_cons	-42.66418	48.19897	-0.89	0.376	-137.1324	51.80406
/sigma_u	.0149996	.0132617			.0026516	.0848513
/sigma_e	.0558448	.0053291			.0463185	.0673303
rho	.0672887	.1160324			.0005925	.5986586

```
Likelihood-ratio test of sigma_u=0: chibar2(01)=     0.45 Prob>=chibar2 = 0.250
```

10. Appendix D: Evidence relating to individual and aggregate benchmarks

10.1. Outcomes performance on PR14 comparative measures

We have examined water companies' performance across the comparative outcomes from PR14. First, we have looked at which company is the frontier (highest performing) company for each outcome. This suggests heterogeneity in performance. Across the five outcomes, the same company is the strongest performer for two outcomes.

Table 14: PR14 comparative outcomes and Frontier Companies

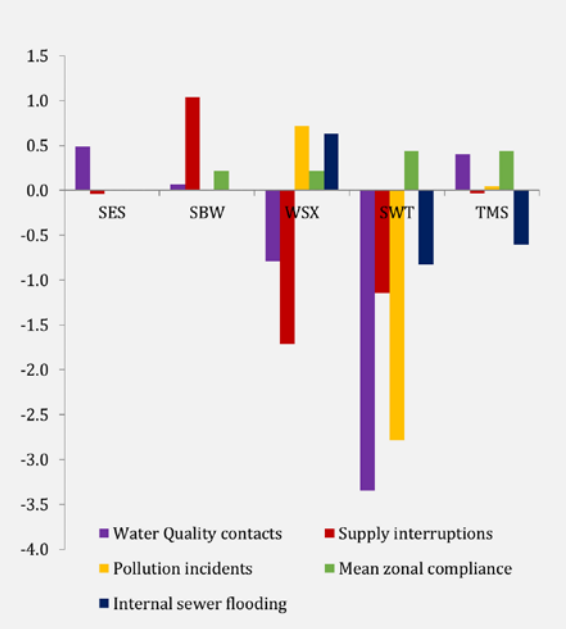
Comparative Outcome	Frontier Company
Water Quality contacts	SES
Supply interruptions	SBW
Pollution incidents	WSX
Mean zonal compliance	SWT & TMS
Internal sewer flooding	WSX

Source: Economic Insight analysis of Ofwat comparative outcome data

Of course, if there were only small differences between the highest performers for each outcome, different companies could be at each frontier, but heterogeneity in outcomes would still, in practice, be limited. We have therefore also examined how each of these frontier companies performed across the other outcomes. Specifically, we calculated *the number of standard deviations* that each company's outcome measure was from the upper quartile outcome, with positive values denoting an outcome *in excess* of the upper quartile.

The figure overleaf shows this information. This demonstrates that, although some companies maintain a reasonably high performance across the outcomes, other companies (that were the frontier for one outcome measure) *performed significantly worse on others.*

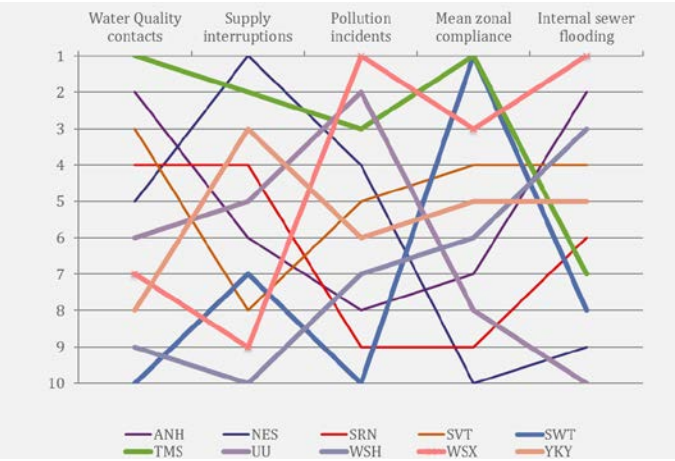
Figure 25: Performance of frontier companies across all comparative outcomes (standard deviations from upper quartile)



Source: Economic Insight analysis of Ofwat 'upper quartile comparative assessments' data

This heterogeneity is also true of firms' performance more generally. SWT, in particular, was the joint highest performer on mean zonal compliance, but the lowest performing WaSC for both water quality contacts and pollution incidents. The figure below shows WaSCs' rankings relative to all other WaSCs for the comparative outcome areas. If firms' performance was consistent across outcomes, we would expect limited crossing of the lines in the figure.

Figure 26: WaSC's rankings across comparative outcome areas



Source: Economic Insight analysis of Ofwat 'upper quartile comparative assessments' data

Overall, this evidence suggests that there is significant heterogeneity in firms' performance across Ofwat's comparative outcomes – consistent with there being trade-offs across them. Firms that are frontier performers in some outcomes areas are significantly behind the frontier in others.

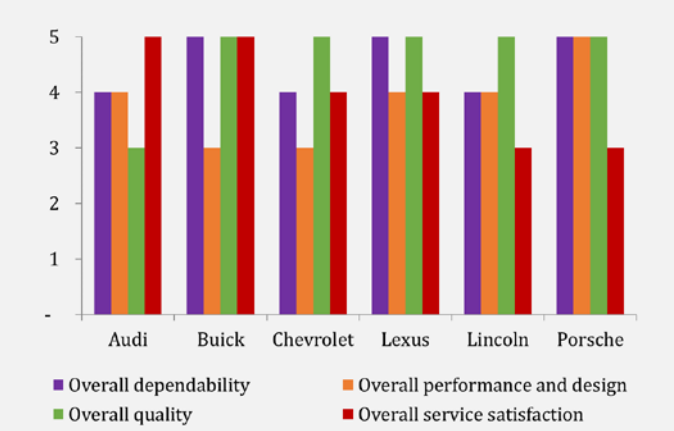
To further illustrate the above points, we have examined outcome performance measures in two competitive markets: cars and airlines. Here it is important to note that the economics characteristics of these industries is quite different from that of water. Therefore, this evidence provided is more to examine whether heterogeneity in outcomes performances *also exists in markets considered to be competitive.* This is to help address any concern that the heterogeneity observed in relation to water might itself be an artefact of 'inefficiency'.

10.2. Cars

Our analysis uses outcomes as measured by the JD Power Survey. This provides information on car manufacturers' performance across four main areas: (i) customer service; (ii) dependability; (iii) quality; and (iv) performance and design. Within each of these four categories, there are a range of sub-measures. These are based on ratings of 2 to 5. A score of 5 denotes "among the best", 4 is "better than most", 3 is "about average" and 2 is "the rest".

We have examined data to determine which companies were among the top performers. Six companies' averaged a score across the four areas of four or more. These companies' scores across the performance categories are shown in the figure below. Despite being among the highest performers overall, all but one company was rated as "about average" in one category. Two were rated as average for overall performance and design, two for overall service satisfaction, and one for overall quality. Put simply, *even firms that perform strongly overall do not achieve the highest outcome scores across multiple outcome areas simultaneously.*

Figure 27: Performance of highest average firms across performance categories



Source: JD Power car ratings

We also examined variations in scores for sub-measures within the above categories for individual manufacturers. This is also consistent with there being variation in outcomes performance at a 'granular/ individual' outcome level. As with the water sector evidence, this might suggest trade-offs exist in outcome performance.

In particular, to get a broader sense of the extent to which even high-scoring firms fall behind the 'frontier' performer, we have calculated, for each of the six companies shown above, the number of sub-categories for which their score was lower than the highest performing firm. This shows that the highest performing firms overall were behind the top performer in between four and eight out of 20 sub-categories.

Table 15: Number of sub-categories for which firms are behind highest-scoring company

Make	No of sub-categories for which the firm is behind the frontier (out of 20)
Audi	4
Buick	8
Chevrolet	8
Lexus	4
Lincoln	7
Porsche	8

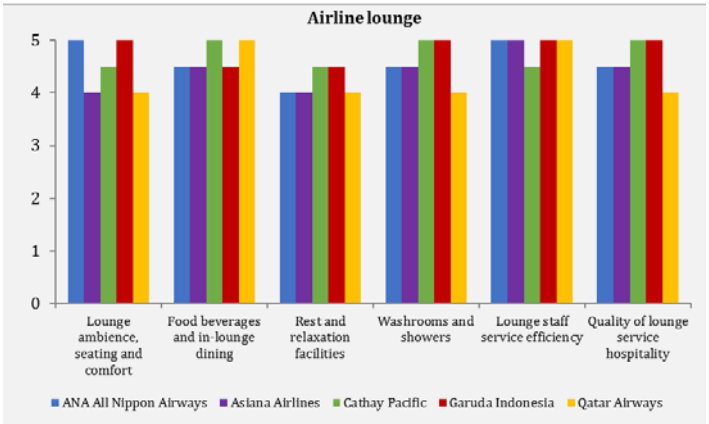
Source: Economic Insight analysis of JD Power car ratings

10.3. Airlines

Skytrax provides a global star rating system for airlines. This provides overall quality ratings, alongside ratings for a range of performance categories, including: the on-board product; cabin staff and airport service (e.g. lounge).

We have analysed performance ratings for five airlines that were rated as "five star" overall. In total, eight airlines had such a rating, but the categories on which the remaining three airlines were rated were slightly different. Again, we find that even firms with the highest overall rating are not 'frontier' across outcome areas simultaneously. To demonstrate this, the figure below shows airline lounge ratings for the five airlines across a range of sub-measures.

Figure 28: Airline lounge ratings for high performing airlines



Source: Skytrax ratings

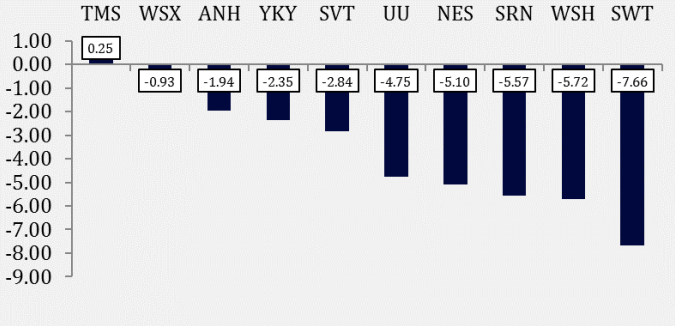
In summary, the evidence in relation to airlines also demonstrates heterogeneity in outcomes performance across outcomes categories, even for firms that have 'frontier' scores for outcomes overall.

10.4. Illustrating overall distance from aggregate frontier using standard deviations

To illustrate how an aggregate benchmark could be constructed for WaSC, we have calculated a candidate average benchmark using firms' standard deviations from the upper quartile performance across each of the common performance commitments subject to comparative assessment. This gives a standardised measure of each firm's distance from the benchmark, which: (i) does not depend on the units in which the particular commitment is measured; and (ii) takes into account observed variation.

There are a number of ways that one could approach combining these standard deviations into an aggregate measure. For example, one could weight each outcome by the amount of revenue that is at risk under firms' outcome delivery incentives. Here, we took the approach of attaching equal weight to each outcome, calculating the aggregate benchmark as the sum of standard deviations. (We could equally have calculated this using the *average* number of standard deviations, in which case the scores in the figure below would all be scaled down by a factor of 5.).

Figure 29: Aggregate benchmark for WaSCs calculated as the sum of standard deviations away from the upper quartile for each comparative outcome



Source: Economic Insight calculations using Ofwat data

As the benchmark for any one performance commitment is defined as the upper quartile, most firms are *behind* the 'overall' benchmark and so record negative scores. This is unsurprising, given the observed heterogeneity in firms' performance across the various outcomes. A negative score does not, therefore, indicate *poor* performance. *Differences* in firms' scores are, however, more meaningful. That is to say, the aggregate measure not only provides a ranking, but the extent of the 'difference' provides an indication of overall 'distance' from the benchmark. Intuitively, however, we would consider 'weighted' approaches (reflecting value at risk across individual outcome areas) to perhaps be superior. This is an area, therefore, that Ofwat should explore further, were an aggregate approach to be considered for PR19.

Economic Insight Limited

125 Old Board Street
London
EC2N 1AR
www.economic-insight.com