Ofwat ODI Research Peer Review

Synthesis Report

Anglian Water, Northumbrian Water, South West Water and Wessex Water

April 2022

4 City Road
London EC1Y 2AA

+44 (0) 20 7580 5383
eftec@eftec.co.uk
eftec.co.uk

This document has been prepared for Anglian Water, Northumbrian Water, South West Water and Wessex Water by:

Economics for the Environment Consultancy Ltd (eftec)
4 City Road
London
EC1Y 2AA
www.eftec.co.uk

Study team

Dr Silvia Ferrini, University of East Anglia
Prof. Nick Hanley, University of Glasgow
Allan Provins, eftec
Dr Scott Reid, ICS Consulting

Reviewer

Ece Ozdemiroglu, eftec

Disclaimer

Document evolution

Peer reviewer instructions	04/03/2022	Ece Ozdemiroglu
Compilation of reviews	23/03/2022	Ece Ozdemiroglu
Compilation of reviews	08/04/2022	Ece Ozdemiroglu
Final Draft report	25/04/2022	Ece Ozdemiroglu

This report is based on eftec's Version 3 – January 2021 report template.

Contents

1. Requirement

1.1 Purpose

eftec has been commissioned by Anglian Water, Northumbrian Water, South West Water and Wessex Water to review the Outcome Delivery Incentive (ODI) customer research by Ofwat. The purpose of the review is to aid the companies' understanding of the Ofwat research in order to: (a) help them provide constructive challenge on the research approach; and (b) assess its implications for their customer research for PR24 and longer-term planning.

The review is based on the following documentation that was made available:

- Ofwat (2021) Invitation to tender (ITT) for Outcome Delivery Incentives (ODI) research: design of methodology, updated 29 September 2021.
- Accent and PJM Economics (2022) Outcome Delivery Incentive Research: Design of Methodology, Stage 1 report, January 2022.
- Outcome Delivery Incentive Research Draft Questionnaire, Accent v14, dated 31st January 2022.

1.2 Overview

The Stage 1 report for the ODI research proposes a novel approach that is (virtually) un-tested in the sector and – to our knowledge - beyond. The research is intended to be used for setting of ODI rates. The report also presents an expectation that the research findings will also be used by companies in their PR24 investment planning processes. Given this, it is important for companies to have a full view on the research methodology and the results it will produce, and to understand how these fit with both the short-term performance and long-term strategic planning needs that PR24 Business Plans will address.

Future challenges from climate change and improving environmental performance have been placed at the forefront for PR24. These can be contrasted with the maintaining and improving day-to-day service levels, determining the appropriate pace of investment to address long-term objectives – for example related to water resources and resilience - whilst ensuring the affordability of customer bills for both current and future customers. Identifying the right path to achieve sustainable outcomes requires addressing various trade-offs associated with different likelihoods of future outcomes. These could be between the short-term goals that PR24 performance commitments and ODIs will be attached to, but also between short- and longer-term goals for the sector.

The views of customers have been a key input for companies in identifying investment priorities and the pace of investments in water and wastewater services, improve their resilience to climate change, and contribute to a healthier natural environment. How customers trade-off between current and future service levels and value the likelihood of different service outcomes is the kind of information companies need in this process. Historic service levels and customers' expectations about, and entitlement to, their maintenance is not sufficient given the scale of changes expected in future. In fact, the UK Climate Change Committee recommended incorporating climate change scenarios into business planning (CCC, 2021).

The water sector has the advantage of having decades of experience of carrying out customer preference research. Good practice research has been based on economic principles, using methods that are consistent with demand theory. Over the years these methods and their application have continued to improve. They will continue to do so – the key challenge is to understand which methods are appropriate for which specific decision-making context.

In summary, companies need customer preference information to better understand a larger set of risk-based outcomes when planning their future investments and service delivery, before the outcomes materialise ("ex-ante planning"). This is a different, albeit related decision context, to the value of an ODI, which is about what to do when faced with a certain outcome of under or over-performance during delivery ("ex-post incentive or penalty setting").

2. General comments

We summarise our general comments under three headings: (i) Appropriateness of the approach for the evidence needs of Ofwat and the companies (Section 2.1); (ii) Design and implementation of the proposed methodology (Section 2.2); and (iii) What could be done instead (Section 2.3).

2.1 Appropriateness of the approach for the evidence needs of Ofwat and the companies

Background understanding

Ofwat could use any criteria to set ODI rates for under/over-performance, such as the cost of delivering the service or making up the shortfall, or even on non-economic punitive or reward grounds. It could also choose to apply weightings to ODI rates that reflect differences between company regions influenced by location or historic factors. To date, though, Ofwat has consistently explained that ODI rates will be set in accordance with marginal benefit values for each performance commitment area, and that rates will be "informed" by customer research. The conventional interpretation is then of an economic value framing for ODI rates, calculating marginal benefits from a demand perspective (and presumably reflecting the full social welfare impacts of doing better/worse than a performance commitment target level).

Conceptual overview

The method and design proposed for the ODI research is a simple one. Essentially, it seeks to elicit preferences for a small set of certain outcomes and extrapolate these to other scenarios. This is in contrast to the present approach in the sector that presents the risk of outcomes happening, which is the most relevant decision context at the time of business planning around potential changes to future levels of service.

The proposed approach, however, is un-tested. More importantly, it is not clear from the information available how the design could create a demand function and value trade-offs against a clear reference point (i.e. a common status quo representing current service levels). The authors refer to 'expected utility' for extrapolating values for certain outcomes but its use is not clear to us given that non-linearity of preferences (along the level of service and level of risk) is not acknowledged.

Therefore, it is not clear that the results will be consistent with: (a) Ofwat's requirements for estimating marginal benefit values; or (b) companies' use of customer preference evidence for investment planning in a cost-benefit analysis (CBA) framing.

Given the novelty of the approach that is proposed, it would have been helpful to see a formal description of how the approach will be consistent with conventional demand theory and some field tests of novel aspects of the approach. This requirement is not specific to this ODI research: in our view, any improvement needs to demonstrate adherence to demand theory and be shown to provide a robust measurement of preferences (i.e. avoiding/controlling for potential biases). Previously, the emphasis for improvements in methods has been placed on presenting an appropriate level of contextual information to inform respondents (for example including comparative performance information), as well as being more engaging for participants (to ensure credibility and the considered and reliable responses are collected from customers).

The proposed design for the ODI research takes a unique approach, effectively asserting that context and the factors that influence demand, such as current performance, the target or proposed change in service, as well location specific issues are not essential to estimating marginal benefit values. The authors find the conventional Dichotomous Choice Experiment (DCE) approach that provides such information problematic, but the problems identified are not well evidenced (providing selected references to poor quality designs). Overall, the assumptions made about context-free valuation[1] and the linearity of preferences[2] are not only untested but are generally counter to the evidence to date.

In summary, we cannot see how the proposed design could produce a demand curve. It does show indifference curves – the trade-off between two impacts. This may suffice for setting ODIs but it is not sufficient for business planning. The absence of a status quo in the design means we have no idea where on the demand curve marginal benefit values are estimated. This gap in the information provided means respondents may be valuing these certain outcomes from their own experience point or from no reference point at all. This is the biggest weakness of the approach. Resulting values are likely to be abstract. It is debatable that they could be reliably compared between company regions by Ofwat – i.e. how could a "high" or "low" value be reconciled if no reference point is given? And values could not be reliably aggregated over a customer base to measure the total benefit of improved / avoided deterioration in service for use in investment planning by companies in the absence of a reference point.

Specific point about complexity and its presentation in customer research

The authors' argument that presenting context-specific information – such as the current level of service - makes surveys too complex is not supported by evidence. Context-specificity has been shown to influence preferences in majority of the studies. The authors' assertion that goes against this evidence should be more thoroughly tested. The examples quoted by the authors (such as Figure 3 and 4 in the Stage 1 report) are examples of poor application, not of poor technique.

Given the proposed departure from the conventional approach to estimating demand, a reasonable expectation for the research methodology would be for it to produce evidence from pilot surveys that

[1] For example, respondents value water cuts the same regardless of how well or badly other services are delivered

[2] For example, if respondents value a 6-hour water cut at £50, they are assumed to value 24 hours at £200 even if they may be indifferent to the impact after a number of hours and hence value 24 hours at less than £200.

compared different designs (theory compliant DCE and the proposed study design) to provide some empirical justification, but such evidence is missing.

Moreover, approaches such as evidence-based policy in the public sector, and Taskforce on Nature-related Financial Disclosures (TNFD) (beta version - March 2022), mainly for businesses, recommend gathering and integrating location and context specific information into decision making. Thus, methodologically and strategically, it is not clear why context is ignored.

Triangulation

It could be argued that setting financial incentives / penalties in response to certain outcomes from delivery should require more precision than evidence that informs on the balance of costs and benefits for Business Cases and value-for-money type assessments. Using the results of a single study with a novel methodology – especially given the lack of testing - cannot offer sufficient rigour for setting ODI rates. Even for investment appraisals that input to business planning, which can usually accommodate wider error bands, Ofwat encouraged companies to develop an evidence base that drew on multiple evidence sources and triangulate these. Unless a subsequent triangulation exercise is planned by Ofwat – including how to deal with values that are measured against reference points and those that are not – the approach being taken is counter to Ofwat's own expectations and guidance.

How does this help companies?

Company business planning involves understanding and comparing the likelihood of delivering different levels of company services. This requirement is increasingly important in the context of developing both long-term delivery strategies and their PR24 business plans.

If the companies are to benefit from customers' views, customers need to be presented with the realities of business planning which include small changes in small risks to service delivery, but also changes to environmental quality, social impact and the resilience of all of these to future changes. Such complexity is possible to convey if research is designed according to economic theory and satisfactorily tested.

Since the decision context of setting ODIs and making business plans are different, a simpler approach focused on certain outcomes that may suffice for the former will not be appropriate for the latter. Moreover, companies use customer preference evidence on improvements and deteriorations in service levels including issues like resilience, environmental value and related trade-offs to be informative for investment planning, particularly in relation to the long-term objectives for the sector. The ODI research seems to be focused on losses.

In conclusion, while we cannot link the ODI approach clearly to demand approach, a simpler approach of valuing a certain outcome could, in principle, be relevant. However, for practical applications, our design- and implementation- related comments below should be taken into account.

2.2 Design and implementation of the methodology

The authors assume that previous DCEs were too complex for respondents leading to variation in preference measurements and scope insensitivities which they see as problems that can be addressed by the design they created.

As these points relate to survey design, we cover them here:

- **The variation between company results is too large.** There is no benchmark for what constitutes a variation that is 'too large'. While there could be an order of magnitude difference between company results, there has not been a comprehensive study (e.g. a full meta-analysis study using the *actual* data collected) to explain such variation as local context, overall company performance, differences in population, as well as survey design and data analysis elements.
- **People cannot deal with the level of complexity created by presenting contextual information.** People face complex trade-offs (like searching for the best insurance, mortgage product) all the time, yet they still make choices. Poor applications of methods that present such trade-offs to individuals do not mean that the methods are poor. It is perfectly possible to design DCE exercises that people can answer in a meaningful way from a welfare economics perspective and from which reliable WTP values can be inferred. It implies a need to think carefully about the overall choice problem set for the respondents, how to describe the attributes, and which levels to pick for these attributes. Even a cursory glance at the DCE studies conducted since PR09 would show various differing ways that attributes and outcomes have been described to customers, ranging from a focus on the private benefit to customers (a small change in a small risk level) to describing public goods and non-use benefits that reflect the full total economic value (TEV) of the outcomes delivered (or not) by companies (e.g. total number of customers that may be affected by a change).
- **People are insensitive to scope of changes.** If we don't observe much/any sensitivity to the scope or scale of an attribute (how far along the value function we move), then this can have multiple interpretations. First, we have chosen the wrong measure of definition to measure scope – for example focusing on small changes in small risks Second, the valuation study is poorly designed. Third, people are genuinely indifferent to the scope of the attribute over the range included in the design – for example if service levels are high and experience of service failures is low. None of these imply that DCE is the wrong approach.

There are other, more specific, design elements of the research that would benefit from attention by the authors:

- **Willingness to Accept Compensation (WTA) format.** There are two issues here: (1) what is the correct welfare measure? And (2) if WTA is the correct measure, can we estimate it in a demand-revealing way? The answer to (1) is that there is no uniquely correct choice, it depends on what we understand about property rights (de facto or de jure). People can be said to have the right not be to subject to sewage pollution in their homes but what about other services? And what about right to zero pollution if the implication is an increase in other risk factors or risky outcomes?

  With regards to point (2), there is no incentive-compatible way to ask a WTA question. And using WTA introduces a risk that people will find the hypothetical market non-credible (e.g. why would a water company compensate a customer for not raising water quality in their local river?).

  More importantly, if ODIs are applied to both under-performance (bill decrease) and over-performance (bill increase), respondents should be informed about both possibilities. The lack of explanation in the questionnaire about what will happen to customer bills in the context of ODI incentives or penalties is a major short-coming in the content validity of the exercise.
- **Coverage of Total Economic Value.** The design does not capture non-use values for environmental or altruistic outcomes (in fact TEV in general). Respondents are instructed to exclude altruism but include preferences for local environment which is contradictory. Values will be partial and not reflect the full social

welfare change that a business planning or cost benefit analysis approach would entail.

- **Contextualising choices.** In a standard DCE, respondents value (the marginal utility of) each attribute in a standard choice task alongside other attributes as well as relative to marginal utility of money. This means the trade-off is against money *and* other service measures simultaneously so that we can capture, more fully, things like substitution effects (or at least we could in principle). This is lost in the compensation task and the impact exercise is limited (as money not involved). This means it does not reflect well the context that underlies investment prioritisation – i.e. understanding what delivers the outcome when weighing up alternative investment options in different service areas.
- **Sampling.** Sample size is generally small on a per company basis given the typical desire to segment results across a variety of dimensions and does not seem to be quoting for vulnerable hard-to-engage groups in each company area. Sampling method of sending people mail to invite them to get online is cumbersome and could pose further risk to acquiring sufficient, representative samples. The approach to weight the responses from the survey is a black box at the moment and variations such as across companies and across water and wastewater companies are not known.

With regards to implementation, there are two key comments:

- It is not clear how the study is planning to produce data from representative samples from each company area (and from different groups in the society – including hard to engage ones).
- Finally, from a purely logistical perspective, if data is made available in the Summer 2022 and ODI rates in December, this will be too late for the companies – especially given the uncertainties about the adequacy of sample size and coverage of different types of customers in each company area.

2.3 What could be done instead?

The first option would be **not to base ODI rates on marginal benefit values**. Approaching this from a regulatory economics perspective, penalties and incentives that are linked to the costs of delivery, clean-up etc. would be a simpler way of dealing with this challenge. Alternatively, rewards and penalties could be set by qualitative research / engagement with customers.

Second, Ofwat could try to find ways of **making use of all the past evidence** from the water companies. Collating the respondent data (not just the results) from past surveys could be time consuming, and hence cost. But there is such wealth of customer demand information in the company data that it is inefficient to disregard that entirely.

Third, it would be possible to **design a DCE** to elicit customers' preferences for a common set of attributes and levels that that was compliant with economic theory, presented a status quo zero additional cost option, with, contextualised service level changes and which uses an incentive compatible payment vehicle.

Finally, if the objective is to avoid DCE, other methods like best-worst scaling that was used in PR19 could be tested – but with the proviso that an approach would need to be compatible with reliable demand estimation. And if the objective remained to **use the proposed approach in the Ofwat research**, it should be **tested more widely** before committing to it: in different places and contexts and against other approaches to learn about the pros and cons more.

3. Specific questions

The points raised in in Section 2 above are further detailed under specific questions with regards to: (i) the design of the Ofwat research (Section 3.1); (ii) its ability to deliver customer evidence needed (Section 3.2); and (iii) comparison of the two and implications for water companies (Section 3.3).

3.1 Ofwat research

a. What is the Ofwat research methodology?

We could describe it as a two-step stated preference exercise: (i) what the authors call 'impact-based analysis' showing relative utility / importance weights for service losses across a set of attributes and (ii) money / utility trade-offs for two attributes (acting as pivots). The idea is that the repeated impact analysis reveals the relative preferences of individuals for different packages of service attributes, whilst the second half (a contingent valuation design) provides a WTA value around which to calibrate these relative preferences.

The intention is to estimate the marginal benefits of desirable attributes and decreases in undesirable attributes. A WTA approach is used as the basis.

b. What are its strengths and weaknesses?

Strengths are that the design seems a cognitively straightforward set of tasks for respondents relying on relating each Performance Commitment (PC) to direct impacts on customers. This design is motivated by a view that the generally small risk changes experienced by customers are not easy to understand and value. But this is asserted rather than demonstrated (through, for example, any pilot evidence). We believe the approach overstates the case for change.

Avoiding the need to specify service levels as part of the design is presented as a benefit of the approach. But this also moves the design away from the realism of actual water company planning, which is largely around (relatively) small risk changes. Furthermore, a great deal of weight is placed on the scope insensitivity argument which is not well articulated or explained.

On the **weaknesses**, our main concerns are that:

- We have no idea what the demand-revealing properties of this type of experimental design are – in contrast, for standard DCE and contingent valuation exercises, we know a lot about how to ask demand-revealing questions. What this means is we have no idea how close to true, underlying WTP the estimated values will be;
- The proposed design does not attempt to measure perceived outcome or payment consequentiality, which has been shown to be very important to the use of stated preference methods in revealing underlying true values;
- Values estimated in a WTA framing will be systematically different to those estimated previously in a WTP framing;
- The proposed design introduces compensation for service delivery decline rather than payment of higher water bills for service improvement. It is not clear to what extent people would find this credible;

- Preferences are assumed to be linear – there is no point at which people will be indifferent to a decline in service delivery (e.g. an 18 hr rota cut may be valued the same as 24 rota cut but very differently if this is the only failing service or one of many); and
- Information on status quo is missing and compensation assessment implies a decline of the current situation vis-à-vis a compensation. Scarpa et al. (2005) clearly demonstrate the profound status quo bias in water choices and the design of the Ofwat study would (all else equal) likely result in an overestimate.

c. How does it contrast to the established customer preference research approach from PR09 to PR19?

Both approaches are based on customer preferences. Beyond the question about the demand revealing properties of the proposed approach, the biggest contrast is the use of compensation as the payment vehicle. This moves away from the established methods of using bill impacts and introduces an untested approach which risks being inaccurate and imprecise (in terms of how it is described to respondents). This gives rise to concern about payment vehicle bias.

Overall, the case for change that is articulated feels weak. The evidence presented is not convincing that DCE approach needs a re-think. It is simply assumed that all previous DCE work has been too complex and leads to unreliable value estimates due to scope insensitivity.

We think the report makes unfair criticisms of the DCE approach (possibly combined with a package-level CV) as a way of solving the research problem. There are two key aspects here. First, the authors claim that DCEs impose cognitive burdens on respondents which are too high in this context, and lead people to use heuristics to make choices. This implies that we either need to explicitly model these heuristics (such as attribute non-attendance), or else cannot interpret parameters from the choice model as showing how people trade-off these attributes in terms of utility. However, the examples provided (such as Figure 3 and Figure 4) reflect poor design on the part of their authors. Just because someone implements a technique poorly does not mean it is a bad technique.

As summarise previously, it is perfectly possible to design DCE exercises which people are able to answer in a meaningful way from a welfare economics perspective and from which reliable WTP values can be inferred. It is not necessarily a simple research task, and it implies a need to think carefully about the overall choice problem you are setting for respondents, how to describe the attributes, and which levels to pick for these attributes. Asking people to make choices which involve trade-offs against different small risk reductions is conceptually demanding, but not insurmountable. This abundantly obvious from the use of DCE is many sectors and many policy-contexts and there is no reason to belief that the water sector is a special case. For example, the estimation of the value of statistical life or a life year lost – which involves changes in small probabilities - is routinely estimated from stated preference research using DCE designs. These are accepted and applied as standard values in Central Government appraisal.

If we observe from a particular DCE that choice model parameter estimates do not show much/any sensitivity to the scope or scale of an attribute (how far along the value function we move), then this can have multiple interpretation. First, we have chosen the wrong measure of definition to measure scope. Second, the valuation study is poorly designed. Third, people are genuinely indifferent to the scope of the attribute over the range included in the design. None of these imply that DCE is the wrong approach.

One reason the authors give for moving away from the DCE approach is so that the choices can be displayed well on a mobile phone screen. Multiple diverse surveys are daily conducted online and evidence of space on the screen being a limitation has not occurred to us (Evans et al., 2018).

d. Is the focus on willingness to accept compensation (WTA) appropriate given the context for PR24 and the longer term challenges faced by the sector?

There is nothing wrong with choosing WTA as a welfare measure (a money metric of the underlying change in utility) even if it is not the most preferred approach. As the report makes clear, both WTP and WTA are candidate money-metrics for the same environmental quality change. We know that, due to loss aversion and endowment effects, WTA values tend to be larger than WTP values for the same underlying quality/quantity change. This is also the case in water sector as shown by Scarpa et al. (2005) and Scarpa et al. (2007).

More importantly, though, whilst the literature is clear about how best to design stated preference exercises to reveal underlying demand in a WTP context, it is much less clear in a WTA context. In particular, economists have worried about the incentive compatibility of WTA formats in stated preferences (Lloyd-Smith and Adamowicz, 2018). As Vossler et al (2012) have shown, in WTP formats we know that stated bids can closely approximate true WTP for public goods when four conditions are met: (i) participants care about the outcome; (ii) a single binary choice is offered on which participants vote yes or no; (iii) the payment mechanism is coercive, and (iv) the probability that the proposed project is implemented is weakly monotonically increasing with the proportion of yes votes. In the literature, conditions (iii) and (iv) have been addressed in terms of perceptions about outcome and payment consequentiality. However, despite encouraging empirical work by Lloyd-Smith and Adamowicz (2018) showing that for public goods, these conditions can also predict incentive compatibility for WTA formats, no such result was found for private goods. Moreover, the Lloyd-Smith and Adamowicz (2018) result for public goods "suggest that WTA values can be valid as long as responses have consequences for respondents". We do not see anywhere in the proposed design where the authors test for perceived consequentiality.

Moreover, Bush et al. (2013) have shown a strategic incentive exists to overstate values in a WTA context, even with a provision point mechanism in place. Aspects of service delivery in the current ODI context are clearly private goods – I can be excluded from supply, and my consumption of the good competes with others' consumption of the good. What this means is that choosing a WTA format makes it harder to demonstrate that stated values are unbiased estimates of underlying true values.

Given all this, WTA does not seem appropriate for considering the longer term challenges such as greater levels of resilience to climate change, higher levels of environmental performance and the trade-offs between pace and timing of investments versus what is affordable. The focus on WTA works fine when framed in terms of the current endowment of service provision but it is less applicable to the future challenges.

e. Does the methodology approach follow best practice?

The impact exercise follows standard approaches to deriving relative utility weights that are tried and tested in the sector. This is not so for the compensation exercise. Furthermore, the combination of the two approaches does not have strong precedence, which is not a problem in itself, but also is not tested and evidenced within the study. The authors also acknowledge that the method has only been used in water sector to value avoided supply interruptions.

Finally, it is surprising that the report does not even reference what most academics feel is the best recent overview of good practice for stated preference methods (Johnston et al., 2017).

f. Is it likely to produce unbiased estimates of ODI rates that will drive the right behaviours by companies, and as a result efficient outcomes for customers?

The most likely outcome is that the research approach will generate inflated ODI rates as they based on a compensation vehicle. It cannot be assumed that these would be unbiased estimates.

The linearity assumption mentioned in the summary above is also worth repeating here. The impact-based exercise will determine a marginal importance of a single attribute assuming linear effects. Therefore, if the customer values 6 hours water interruption at £50, the linear assumption of £200 for 24h (£50/6 hrs x 4) is a pessimistic estimate that penalises companies, if, customer is actually indifferent after 18 hours. At the same time, if the industry improves a service for 6 hours above the limit, the customer might not be happy to pay £50 as described by the well-known prospect theory problem.

g. What sort of confidence intervals for ODI rate estimates can be expected from the approach? Do other approaches generally produce more accurate estimates?

It is impossible to predict the confidence interval since the final result is a function of two estimates. Surely, in breaking the exercise in small pieces, the statistical error of each task is getting smaller but the final error can be quite large. Specific statistical techniques might be needed to get variance of combine methods (Wu, 1985) but nothing of this kind is mentioned in the report and it is not normal practice for consultancy applications. So, it is difficult to fully compare the improvement of the confidence intervals of the new methods versus the standard DCE.

The question that follows is: what does a regulator do with an ODI rate that has a wide confidence interval? The only answer we can think of is that this is dealt with in the design of the ODI incentives themselves – through the cap and collars (ranges of out-/under-performance which limit the application of the ODIs – meaning that we allow a lot of wiggle room before the ODI rates apply).

h. What are the strengths and weaknesses of the combination of the proposed 'impact-based exercise design' and the 'compensation-based exercise design'?

As mentioned above, the combined approach is not tested and validated. The only obvious benefit seems to be facilitating the display of the questionnaire on a mobile phone screen since the questions just focus on attribute levels and nothing else.

The authors argue that "one would need only to estimate the value for a single avoided incident via the compensation approach. This would then serve as a 'pivot' value against which the estimated impact index could be used to derive the values of every other service issue". There needs to be some testing of how valid this approach of pivoting all monetary value estimates for each attribute off a single compensation scenario. Would the resulting monetary values for each attribute depend on which compensation scenario was chosen as the pivot? In the draft questionnaire, two compensation questions are used – this is a result of the later discussion in the report on the desirability of using two different compensation scenarios to provide two alternative pivot values. But the essential problem still remains - we do not know how sensitive the final value

estimates for each attribute will be to which compensation scenarios are used. How should the "best" scenarios be chosen?

i. Any other aspects of the methodology which remain unclear and requiring further work?

Expected utility theory. It is surprising that the authors introduce expected utility theory as a property of best-worst scaling. Expected utility theory is one way of thinking about how people make risky choices, and can be used as a potential explanation for how people choose in a DCE (e.g. Glenk and Colombo, 2011).

Non-use values. We also question the extent to which non-use values for river improvements (e.g. to low flows) are included in the design. The discussion of existence values around page 45 is a bit dated and also rather misleading. There is considerable evidence that people can hold a significant, separable value for improvements in environmental quality which (i) are not "Local" to them; and (ii) do not depend on their actual use of the resource. In the context of improvements to low flow rivers in England, Hanley et al (2003) found evidence of significant non-use values and a statistically significant distance decay function for such non-use values (although the size of the distance decay effect was bigger for use values than for non-use values). Thus, the survey should consider impacts on river and coastal water quality that are distant to the respondent, and will collect data to identify users from non-users.

Citizen vs consumer values. The discussion of a possible distinction between citizen and consumer values is puzzling. There are many context effects which can make a difference to people's choices, and the framing of those choices is one such context. *How* people form their minimum WTP/maximum WTA is not relevant to the application of cost-benefit analysis; rather, we care about what these amounts are.

Units for monetary values elicited from the compensation exercise. It is not absolutely clear what these are and how they should be interpreted and used in CBA. The description of the payment vehicle suggests that the compensation would be offered as a credit to a customer's water bill (or by sending a cheque – incidentally the reference to the sending of a cheque seems archaic when most customers now pay via direct debit.). We assume that this is a one-off credit and hence a one-off adjustment to the annual water bill. There would need to be further guidance on how to interpret this one-off adjustment when used to determine the annual benefits associated with any investment. This is less easy to interpret compared to a metric like £ increase / decrease in annual water bill each year per customer per unit of service change.

Standard errors. The combined standard error of the two estimates needs to be fully calculated using best practice in statistics and ideally considering the loss in precision vs the reduce complexity/better layout.

3.2 Customer evidence needs

a. What are the customer preference evidence needs of the companies for PR24, particular in relation to long-term challenges around climate, environmental quality and resilience?

PR24 introduces new challenges for customer preference evidence. The focus going forward is much more about the long-term priorities to be delivered and the pace and timing of expenditure/investment to achieve those long-term priorities. Ofwat's emerging framework for PR24 is recognising this shift to long-term priorities; in some respects, it is driving this shift. It is clear that none of these are adequately represented in the ODI research, which is centred more narrowly on service levels in the present.

The longer-term challenges are about sustaining and enhancing what is currently delivered. So, for example, greater levels of resilience to climate change, higher levels of environmental performance and the trade-offs between pace and timing of investments versus what is affordable. A focus on WTA works fine when framed in terms of the current endowment of service provision but is less applicable to the future challenges.

The short- and long-term vision of the PR24 introduces complexity since the customers might know very well which services to prioritise in the short term but might be unsure about the future. The starting point, though, will be to build customer focused descriptions of future scenarios, informed by the available long-term planning information and projections (e.g. target outcomes and constraints). Choice tasks and exercises can be constructed around these scenarios to test the sensitivity of customer preferences – i.e. what are the key factors driving customers choices (bill impact, time to meet targets, the "destination" for environment and resilience of services). Some testing will be needed though, since the research will need careful design to ensure respondent understanding.

b. What do companies need to know to be able to compare investments in delivering short- and long-term priorities?

The marginal benefits of customers seem essential to consider marginal costs to invest in specific improvements. These benefits should reflect the possible investment levels of the company to facilitate a cost-benefit assessment. The need is not solely to be able to compare short- and long- term priorities. Rather companies will need to demonstrate how plans in the shorter term (e.g. the next AMP8 plan) are nested within a longer term strategy (to 2050). This presents new expectations and challenges in terms of providing the customer evidence that support both the longer-term strategy and the shorter term stepping stones.

c. How can companies compare investing in adaptive capacity now for resilient services in the future and avoiding having to deal with putting right the impacts of climate change risks?

Companies can minimize the risks of climate change by considering a portfolio of investments in grey infrastructure, demand management, nature-based solutions or grey infrastructure and seek for each of them the consumers' marginal benefits. These benefit estimate can be used to rank projects and conduct cost-benefit analysis.

To make investment in adaptation truly comparable to other investments, we need to change the way we describe the status quo. Given the realities of climate change, we cannot have a costless status quo. Therefore, our customer research concerning the longer term should no longer present a status quo (no change) that is possible to deliver at no cost. Status quo choice in a DCE exercise can be designed to accommodate a dynamic of increasing cost to sustain service levels, and this is likely to affect subsequent choices as the literature already shows how crucial status quo is in valuing improvements and losses.

The recent experience of the WICS in Scotland and the SRC21 determinations are relevant here. While the regulatory process differs (more akin to negotiated settlements), the question posed here strongly echoes the one that came to dominate the SRC21 process. Through the work of the Customer Forum – to evidence customer views – a consensus emerged that not having to put right the (possible) impacts of climate change risks was worth higher bills and expenditure allowances in the near term. Spend a little now to avoid a lot later (maybe). Effectively Scottish Water was allowed a fund of money they could access as they understand better

climate change impacts. For commentary on this see "Spending for a rainy day: Interview with Alan Sutherland", The Water Report, November 2020.

3.3 Comparison & implications

a. Will the results of the Ofwat research meet the companies' evidence needs?

In terms of overall evidence needs, the short answer is no. The proposed approach will produce individual ODI estimates but these do not reflect the current services of any company. Companies can use these initial estimates to compare their marginal costs and investment in R&D but given high uncertainty around these results are doubtful to support any actual decision on investments.

As noted above a significant new focus will be around long-term priorities and how they are delivered. This has to be integrated with the established and tried and tested programmes of valuation research to support investment planning. The Ofwat research is consciously not designed with this in mind: it is more limited. With respect to the more limited evidence needs around (common) ODI rates, our commentary above raises enough doubt about whether the proposed design will meet even these limited needs.

b. Could the Ofwat methodology be a template for companies to collect their own evidence?

It could. But only within the limited focus noted above. The report explains the approach very clearly in terms of "how to do it". It can give insights on how to prepare visual and descriptive information for ODI and the background information to prepare a valuation questionnaire. The approach needs input from experts in different steps. The freedom to design bespoke questionnaires and analyses is not necessarily clearly specified in the report and companies will have to research how to include specific questions in the proposed questionnaire.

The proposed compensation exercise is too untested for companies. Methodological plurality may create noise in the sense of more, not less, information for a regulator to assess, but it is unwise (at least at this stage) to rely on an untested approach to estimating monetary value for a full suite of water, wastewater and environmental service attributes.

c. How does relying on a single study compare with advice for PR19 of using / triangulating multiple sources of evidence?

This is problematic. It would be hard to compare results from the proposed method with the existing evidence base, because as explained above, we do not know enough/anything at all about the demand revealing and incentive properties of this new approach. The idea to simplify and standardise in a single study seems to go in the opposite direction of triangulation since diversity of views cannot be captured. Standardising ODIs and design setting undermines the companies' business, social and environmental context.

While triangulation was a challenge for many companies at PR19, with the right focus and effort it was delivered successfully by a number of companies. Subsequent to PR19, the updated guidance developed for CCWater (CCWater and SIA Partners, 2021) could help this to improve further. In that context, it is difficult to understand the case for relying on a single study.

d. What are the risks of relying on a single national study in terms of methodology, variety of data and the degree to which location specific factors can be covered?

In principle, a single national study that can capture local and company variations in a full and consistent manner is a good idea. The proposed approach is, however, very risky as it's not tested and not comprehensive. A single national study should focus on crucial aspects (e.g. setting of choice cards, numbers of attributes/levels) and rely on a well-tested approach (e.g. DCE). At the same time, companies should be encouraged to tailor the study to capture local context and company's services.

In any national level study, a research question of interest will always be to test if model parameters and estimated valuations are uniform across regions/companies. As an industry structured largely as regional monopolies and networks, it would be inappropriate to start with any presumption that the "law of one price applies". Rather, our prior expectation would be that preferences and values differ regionally – in the same way that the costs of service differ. With a suitable sample design, restricted functional forms can always be tested but we would always prefer a "general to specific" methodology, which, in this case, would recognise the null hypothesis that preferences and values are not uniform. In Table 8 (Section 4.4) of the Stage 1 report, the proposed sample sizes by company area are in many cases at the *lower* end of what we would recommend to clients and the risk is that the data will only support the reliable estimation of restricted functional forms that impose equality across the regions/companies.

e. What are companies' options to fill the (remaining) gaps?

This is very unclear from the report since the questionnaire, choice cards and analysis seem fixed and rigid. We cannot see how fatigue and learning effects can be avoided if a company wanted to include extra questions. This is another aspect that deserves further research, since the proposed methodology presents limited space for companies' research questions.

In previous publications Ofwat has cautioned companies against duplicating this centralised research (see Ofwat, 2021). The untested nature of various aspects of the proposed approach, with no recognition of a testing phase is notable risk for companies and makes this regulatory expectation mis-placed.

As service providers, it should be sound business practice for companies to understand their customers' priorities for services and how those are valued. The Ofwat ODI research should be seen as a complement to their own efforts, not a substitute.

f. What are the recommendations for Ofwat to improve their methodological approaches and /or the integration of the findings with wider evidence?

We list our alternative approaches in the summary above.

The Ofwat approach is straying away from any previously tested methodology and the key recommendation is to validate it before using it as a model for the sector. Such research (e.g. using an incentivised, experimental approach) should be designed to tell us something about the demand revealing and incentive properties of the proposed method. Such testing could take years so in the meantime Ofwat could consolidate the experience gained in previous Price Review applications and provide overarching guidance on how to better

implement available methodologies (e.g. in relation to DCEs). To gain valuable insights from previous applications, Ofwat can collate previous results (ideally datasets) and enhance this evidence with secondary information (Johnston et al. 2017) to explain the variation of results attributed to local conditions, companies' conditions or consultants' traits.

Another area of research would be to trial the use of the WTA format to see whether (a) people find this to be a credible scenario and (b) whether we get instances whereby a particular service package passes the CBA test when a WTA format is used, but where customers are willing to pay the implied increase in bills to fund this in reality.

From a regulatory rather than valuation point of view, the proposed approach is guilty of putting the cart before the horse. ODIs are regulatory mechanisms designed to motivate the service performance and delivery by water companies. The first focus (the horse) should be on what are the right performance levels. Ofwat unlike other regulators have more recently moved away from the use of economic levels of service and CBA to set PC levels (replacing this with more emphasis and focus on using upper quartile benchmarks).

As noted above, one of our most significant concern relates to the compensation exercise that will be used to estimate monetary values. A preferable option to overcome those concerns would be to combine the impact exercise with a more conventional valuation exercise (that could still be designed as a double bound exercise). For example, we see similarities in the overall design structure to work previously published by Chalak and Metcalfe (2022), which combines a similar impact exercise with a package (multi-attribute) valuation exercise.

References

CCWater and SIA Partners (2021), *Triangulation - A review of its use at PR19 and good practice*.

Chalak, A. and Metcalfe, P. (2022) Valuing water and wastewater service improvements via impact-weighted numbers of service failures, *Journal of Environmental Economics and Policy*. DOI: 10.1080/21606544.2021.2023052.

Evans, J. R., & Mathur, A. (2018). The value of online surveys: A look back and a look ahead. *Internet Research*.

Glenk, K. and Colombo, S. (2011), How Sure Can You Be? A Framework for Considering Delivery Uncertainty in Benefit Assessments Based on Stated Preference Methods. *Journal of Agricultural Economics*, 62: 25-46

Hanley N, Schlapfer F and Spurgeon J (2003) " Aggregating the benefits of environmental improvements: distance-decay functions for use and non-use values" *Journal of Environmental Management*, 68, 297-304.

Johnston et al. (2017) 'Contemporary Guidance for Stated Preference Studies', Journal of the Association of Environmental and Resource Economists 2017 4:2, 319-405.

Lloyd-Smith, P. and Adamowicz, W. (2018). Can stated measures of willingness-to-accept be valid? Evidence from laboratory experiments. *Journal of Environmental Economics and Management* 91: 133-149.

Ofwat, (2021) *PR24 and beyond position paper: Collaborative customer research for PR24*.

Scarpa, R., Ferrini, S., & Willis, K. (2005). Performance of error component models for status-quo effects in choice experiments. In *Applications of simulation methods in environmental and resource economics* (pp. 247-273). Springer, Dordrecht.

Scarpa, R., Willis, K. G., & Acutt, M. (2007). Valuing externalities from water supply: Status quo, choice complexity and individual random effects in panel kernel logit analysis of choice experiments. *Journal of Environmental Planning and Management*, *50*(4), 449-466.

Vossler, C. A., Doyon, M., and Rondeau, D. (2012). Truth in consequences: Theory and field evidence on discrete choice experiments. *American Economic Journal: Microeconomics* 4(4): 145-171

eftec
economics for
the environment

4 City Road, London EC1Y 2AA

+44 (0) 20 7580 5383

eftec@eftec.co.uk

eftec.co.uk

@eftecUK