Outcome Delivery Incentive Research: Design of Methodology

Stage 1 report

January 2022

Contact: Paul Metcalfe
E-mail: Paul.metcalfe@pjmeconomics.co.uk
Telephone: +44 (0) 7786 656834

File name: 3524rep08_Stage1Report_v3.docx

Registered in London No. 2231083
Accent Marketing & Research Limited
Registered Address: 30 City Road, London, EC1Y 2AB

Contents

Executive Summary

Introduction

Accent and PJM economics were commissioned jointly by Ofwat and CCW to develop and test a methodology for obtaining the customer evidence needed to support outcome delivery incentive (ODI) rate setting for common performance commitments (PC) at PR24. These are the PCs that all water companies will be measured against over the next price control period. This project was referred to in Ofwat's recent position paper 'PR24 and beyond position paper: Collaborative customer research for PR24' (October 2021).

This is the final report on Stage 1 of the study. Its focus is on setting out the recommended approach to survey design and delivery to take forward to Stage 2. It discusses the reasons for the recommendations on methodology and the pros and cons of different approaches and addresses a range of issues pertaining to the methodology.

Requirements for the methodology

For PR24, although the final methodology for ODI rate setting is yet to be determined, Ofwat has expressed a preference to continue using marginal benefit estimates to inform the rates that are set but without the link to costs (Ofwat, 2021a). This implies a need for a methodology focused on measuring the value of those benefits.

The table below shows the PCs that Ofwat has indicated to us should be the focus of the methodology, bearing in mind that these may change over the coming weeks as discussions continue as part of the Outcomes Working Group.

Table 1: Common PCs for inclusion in the ODI rates research

	Water only	Wastewater only
Customers receiving excellent service everyday	Water supply interruptions Compliance risk index (CRI) Customer contacts about water quality Event risk index (ERI)	Internal sewer flooding External sewer flooding
Environmental outcomes	Leakage PCC (per capita consumption) Business demand	Pollution incidents Discharge compliance Storm overflows Environmental Performance Assessment Bathing water quality River water quality

Source: Ofwat (2021d)

In line with guidance from Ofwat and CCW, the scope of the design has been focused on one core stated preference survey only, plus guidance on the further research that is required to provide a fuller picture of customer views.

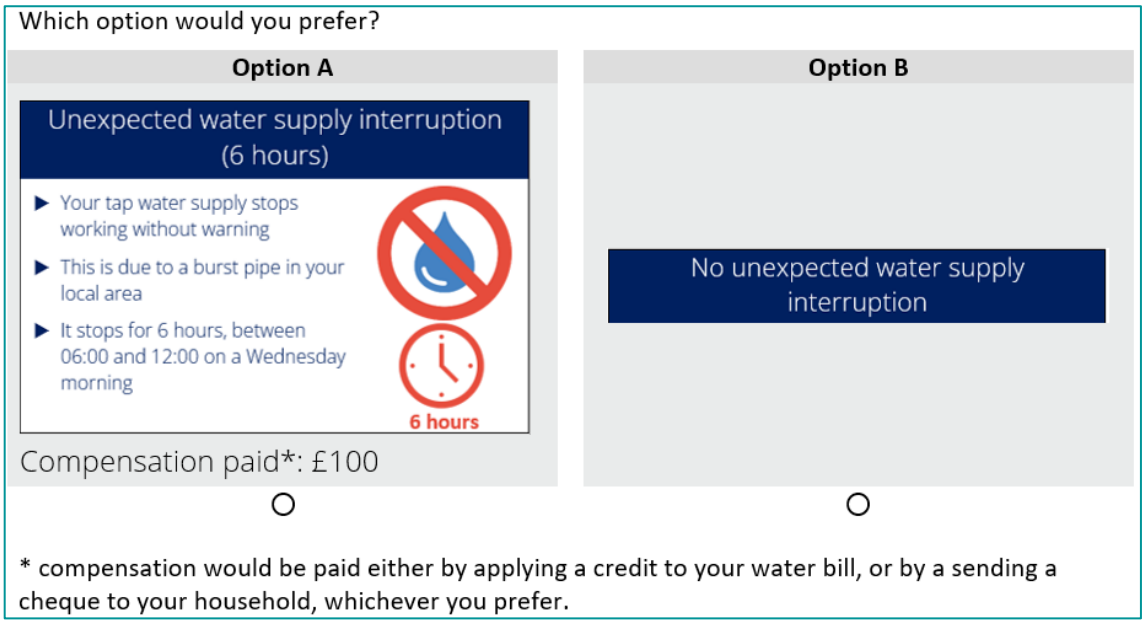
Survey design

Based on the strengths and weaknesses of the selection of options reviewed, our recommendation for the Collaborative ODI research is that the core survey should be design around two choice exercises:

- An impact-based exercise to estimate the relative impact of all types of service failure. (See Figure 18 for an example choice, as it would appear on a mobile phone screen.)
- A compensation-based choice exercise to estimate the values of two different types of avoided service failure. (See Figure 20 for an example choice, as it would appear on a desktop or laptop screen.)

Figure 1: Example impact-based exercise question (household version, portrait format)

Figure 2: Example compensation-based exercise question (household version, landscape format)



These results would be combined in the analysis to return estimates of the values of all types of service failure. For example, if we know from the compensation-based approach that the value of an avoided 6-hour supply interruption is £50 and, from the impact-based exercise, that having discoloured water for two days is three times as impactful, then the derived value for an avoided discoloured water incident lasting two days is £150. (Section 5 contains a full worked example of how the estimates are intended to be derived.)

The key advantages of this combined approach are:

- It is simple and customer focused, whilst remaining consistent with the requirements for economic valuation evidence needed to support the setting of ODI rates.
- It avoids the need for service levels at all, thereby avoiding the 'denominator effect' altogether and its concomitant issues (see Box 1), plus the practical problems associated with companies needing to provide estimates of base service levels and variations around these within a short space of time.
- If there are changes to PC definitions following completion of the core survey, provided the service issues have been carefully chosen to span the range of impacts that water and wastewater services could have, the results should be flexible enough to provide the valuation evidence needed for the new definitions.
- Choices can be set up to work well on a mobile phone.
- Where additional service issues require valuation for the purposes of bespoke ODIs and / or enhancement cases, the method allows for valuations to be obtained in separate studies without any issues caused by the need for package scaling. (See Section 5.5 for discussion of how company research should be designed to link into this methodology.)

A key feature of the method is that it requires all PCs to be mapped to their direct impacts on customers. It is impossible to include a PC such as leakage reduction, for example, as a continuous measure. Instead, the method requires that there must be a direct impact of some kind, that customer either experience or they do not. In the case of leakage, this requires consideration of how leakage actually does impact the customer, which can be debated. On the other hand, by enforcing the discipline of identifying how a PC impacts on customers, and switching the focus to measuring that impact, it is arguable that the resulting valuation is more meaningful as a consequence.

In Section 3.4 of the report, each group of common PCs is addressed, in turn, focusing on how they may be represented in the survey and be mapped back from that representation to the original PC definitions. In each case, there remain a series of issues to be addressed in order to finalise the survey and to set the values of the parameters needed to complete the mappings back to the PC definitions. A set of questions is accordingly included at the end of each of these sections, which we are putting to stakeholders for support in finalising the research methodology.

Survey administration and sample design

Household survey

Two approaches to the household survey have been forward for pilot testing:

- A postal address file approach
- A commercial online panel approach

Both approaches have merits, and hence it is recommended that both should be developed and piloted to explore issues of practicality, efficacy and cost and reflecting the strengths of opinion amongst supporters of both approaches.

Non-household survey

The core approach recommended for the pilot survey is phone-email/post-phone. This involves phoning businesses at random (from, for example, a Dun & Bradstreet sampling frame), then posting/emailing materials, and either carrying on with the interview or booking an appointment to call back to complete the survey.

An alternative approach is being simultaneously developed, which would involve emailing and/or phoning customers via lists obtained from retailers. This could provide potential cost savings though it does mean that two modes of approach – email and phone – are being used which could have comparability concerns. Additionally, the option may not be possible for all retailers which would create clusters of database availability by company given the way that the market is structured. Some smaller retailers may only have higher consumption customers, for example, thus potentially making a representative sample difficult to obtain via this approach.

Discussions are continuing with the retailers to see how many would be able to provide assistance. If the retailer approach develops over the coming weeks, then this could also be included in the pilot testing.

Sample sizes

We are recommending that wastewater and water issues are included together for every customer, rather than having separate surveys for water and wastewater services. This increases the information gained from each participant.

We further recommend that there should be a minimum of 500 households in each water supply area, and in each wastewater supply area, for all companies except for Hafren Dyfrdwy for whom a smaller sample size of 350 is recommended for reasons of proportionality. This will allow up to 3-way segmentations of values within company area.

For non-households, we have applied a global multiplier of 0.4 to all sample cell sizes in order for the total sample to be at least 200 for each water and wastewater supply area. Whilst, based on statistical considerations alone, the sample size for non-households should be at least as large as that for households, a smaller sample size has been recommended due to the fact that non-household interviews are much more costly per interview than household interviews. Accordingly, on an efficiency-per-pound basis, a smaller sample size for non-households than for households is appropriate.

In most cases, the recommended sample sizes for non-households will not allow for any segmentation of values for this customer group within water company area. It will allow robust estimates at a company level, however, for all companies except for Hafren Dyfrdwy. It will also allow for segmentation analysis of different types of non-household at the aggregate level.

For Hafren Dyfrdwy, the achievable non-household sample size for the wastewater service area is likely to be too small to estimate reliable values at even a company level. In this case, we recommend alternative approaches, which should result in improved estimates. (These are set out in Section 4 of the report.)

To the extent that companies wish to pay for a larger sample size, we recommend that they should be free to do so. Moreover, companies should have the flexibility to boost sample sizes in a manner best suited to how they wish to research their customer base. The survey method should be the same for the boost samples as for the main samples, however, for the data to be comparable, and useable as part of the main sample for analysis. Whether or not samples are boosted, weights will need to be derived and applied to ensure that the results are representative for each company.

Analysis and outcomes

Derivation of PC valuations

A rigorous econometric analysis is required to obtain the valuation estimates necessary to derive ODI rates. The steps that are needed are outlined in Section 5 of the report, which includes worked examples.

Further research required

In addition to this analysis, and to the survey methodology set out more broadly, we also recommend that an additional piece of research is undertaken at a late stage in the business planning process to measure customer preferences with regard to the overall relationship between bills and service levels, alongside measurement of the acceptability and affordability of the business plan. This piece of research would serve to set global limits on the degree to which bills could increase during the forthcoming price control period, including for both base service levels and for the bill variation attributable to ODIs. It would not be limited to common PCs only, but would, instead, cover the entire business plan. (More details on this are set out in Section 5.4 of this report.)

Links to company research

The outcome from the Collaborative ODI research will be a set of valuations for a suite of common PCs. Ofwat has already strongly encouraged companies not to submit valuation evidence with a view to challenging ODI rates (Ofwat, 2021b), but expects that companies will potentially undertake further valuation research to support enhancement cases and/or bespoke ODI rates. In Section 5.4 of the report, we set out how companies' own research might best link into the outputs from the Collaborative ODI research to support these areas.

Next steps

The next stage of the study involves testing and refining the methodology. The final output from the study will include a complete set of well-tested research materials, with accompanying experimental designs and supporting documentation, for obtaining the evidence needed to support the setting of ODI rates for common PCs at PR24.

1 Introduction

1.1 Background

Accent and PJM economics were commissioned jointly by Ofwat and CCW to develop and test a methodology for obtaining the customer evidence needed to support outcome delivery incentive (ODI) rate setting for common performance commitments (PC) at PR24. This project was referred to in Ofwat's recent position paper 'PR24 and beyond position paper: Collaborative customer research for PR24' (October 2021).

In commissioning a collaborative national study, Ofwat and CCW aimed to ensure comparability of results by applying a common methodology, and thereby identify genuine differences. Moreover, the study was believed to provide an opportunity to undertake an in-depth review of potential options and identify a best-practice approach that results in estimates that are high quality and fit for purpose.

The project has been structured as a two-stage study. The first stage is a review of methodology options to deliver research to inform ODIs, and the development of a preferred option/s for PR24 ODI research. The second stage is the development and testing of materials for use in this research. The delivery of the customer fieldwork constitutes a third stage, which does not form part of this commission.

1.2 Objectives and scope

The requirements for Stage 1, which is the focus of this report, included the following:

- Consider the range of requirements for the methodology.
- Review relevant research undertaken to date in the water sector and other sectors.
- Develop methodology options that can deliver robust findings.
- Work with Ofwat, CCW and others to refine the approaches and to identify the most suitable methodology for use in customer research to set ODI rates.
- Set out a recommended approach to delivering customer insights that can be used to set ODI rates.

1.3 Work to date

To date, the project has delivered:

- An inception report.
- A report on the findings from an industry consultation.
- A desk review of relevant PR19 and RIIO2 materials, plus broader literature providing guidance on best practice on methodologies for non-market valuation.
- An interim report summarising the above, presented to companies at a collaborative research steering group meeting.

- An industry workshop at which methodology options were presented and discussed.

At the same time as this work has been carried out, Ofwat has provided initial guidance on the common PCs it intends to use for PR24 (Ofwat, 2021c), and which of these it requires valuations for as part of the collaborative research study (Ofwat, 2021d). Ofwat has also published a consultation paper on long-term delivery strategies and common reference scenarios (Ofwat, 2021e). These papers have all been taken into account within the present study, in particular with regard to the common PCs for which valuations are required as this is key to the research design.

Additionally, a customer research study commissioned by CCW and focussed on exploring attitudes towards the proposed coverage of common PCs, and ways in which those PCs can best be worded and presented to customers, has been underway throughout the period of the present study. A final report from this research is due to be delivered in the very near future. Although preliminary findings from this study have recently been shared with us, the outcomes from this research have not yet been fully factored into the methodology proposed herein.

1.4 This report

This is the final report on Stage 1 of the study. Its focus is on setting out the recommended approach to survey design and delivery to take forward to Stage 2. It discusses the reasons for the recommendations on methodology and the pros and cons of different approaches and addresses a range of issues pertaining to the methodology.

- Section 2 details the requirements for the methodology, including the parameters set for the study by Ofwat and CCW, and the principles for success.
- Section 3 includes a review of design options and presents our recommended approach.
- Section 4 similarly reviews options and presents recommendations with respect to survey administration and sampling.
- Section 5 sets out our expectations regarding the analysis that ought to be undertaken of the survey data, as a minimum, and the core outcomes that will be derived. This section includes recommendations for further research requirements and discusses links to companies' own research for enhancement cases and bespoke ODIs.
- Finally, Section 6 sets out the next steps for the study.

Additionally, there are two appendices to the report:

- Appendix A contains technical details concerning the general approach to mapping from service issues to common PCs
- Appendix B contains a summary of stakeholder views on survey design as expressed at the industry workshop, held on 13 December 2021.

2 Requirements for the methodology

This section sets out and addresses a number of issues concerning the requirements for the Collaborative ODI research methodology. These include:

- Type of evidence needed
- PCs to be included
- Scope of Collaborative ODI research
- Principles for success

These jointly set the framework for the review of options and recommendations that follow.

2.1 Type of evidence needed

Since PR14, when ODIs were first introduced in the England and Wales water sector, the rates set for outperformance and underperformance against target levels of service have been set following a formula, set by Ofwat, that was designed based on economic principles. The formula was intended to incentivise the optimum level of service based on the marginal costs and marginal benefits associated with performance variations, taking into account the degree of cost sharing implicit within the broader regulatory settlement. (Ofwat, 2013)

For PR24, although the final methodology for ODI rate setting is yet to be determined, Ofwat has expressed an intention to continue using marginal benefit estimates to inform the rates that are set but without the link to costs (Ofwat, 2021a). This implies a need for a methodology focused on measuring the value of those benefits.

In cost-benefit analysis, value is defined using the measures of willingness to pay (WTP) and willingness to accept (WTA).

- WTP is defined as the amount of money that people would give up in exchange for a service improvement, or to avoid a service deterioration.
- WTA is defined as the amount of money that people would need to be compensated in order to accept a service deterioration, or to lose out on a potential service improvement.

For ODI rates, the same conceptual measures of value apply. The measures relate to the general trade-off between money and service level and are not inherently context specific.

The appropriate method of measurement may, however, depend on the form of application. This issue is accordingly discussed in Section 3.

Additionally, there are important links between these measures and the evidence needed to justify cases for enhancement expenditure, and for setting bespoke ODI rates. These issues are discussed in Section 5.

2.2 PCs to be included

The scope of the research, in terms of the common PCs to be included, has not yet been finalised. Ofwat has provided initial guidance on the common PCs it intends to use for PR24 (Ofwat, 2021c), and which of these it requires valuations for as part of the collaborative research study (Ofwat, 2021d).

The table below shows the PCs that Ofwat has indicated to us should be the focus of the methodology, bearing in mind that these may change over the coming weeks as discussions continue as part of the Outcomes Working Group.

Table 2: Common PCs for inclusion in the ODI rates research

	Water only	Wastewater only
Customers receiving excellent service everyday	Water supply interruptions Compliance risk index (CRI) Customer contacts about water quality Event risk index (ERI)	Internal sewer flooding External sewer flooding
Environmental outcomes	Leakage PCC (per capita consumption) Business demand	Pollution incidents Discharge compliance Storm overflows Environmental Performance Assessment Bathing water quality River water quality

Source: Ofwat (2021d)

The measures in the above table are those for which values are needed; they do not necessarily need to be included directly within any customer research materials. The means by which they are valued, including any translations or re-wording needed, are matters for the research design, and are accordingly discussed in Section 3.

Table 2 also includes measures that are overlapping with one another. For example, CRI, ERI and Customer contacts about water quality are all dependent on similar drivers. Consequently, an important issue to be addressed by the methodology is how to avoid double counting of values. This issue is also discussed in Section 3, along with technical details shown in Appendix A.

2.3 Scope of Collaborative ODI research

In line with guidance from Ofwat and CCW, the scope of the design has been focused on one core survey only, plus guidance on the further research that is required to provide a fuller picture of customer views.

Furthermore, whilst there is potentially merit in considering other approaches, particularly where triangulated insight into customer value can be used within the regulatory framework, only stated preference research has the ability on its own to measure the entire suite of common PC values, and to capture the entirety of the value including non-use value as well as use value. Accordingly, a stated preference approach, which allows the design to be tailored to measure these, is the only viable approach for the present research and is the focus of the methodology presented in this report.

2.4 Principles for success

An initial set of principles was established with the project Delivery group at the outset of the study and consulted upon with companies to guide the development of the methodology. These principles were all broadly approved by companies, and include that the study should be:

Collaborative	Developed in consultation with companies and key stakeholders.
Tailored	Designed with a specific focus on the PCs and measures required.
Customer-focused	The survey should be focused on customers, using language and questions that are meaningful and understandable to them, in order to generate meaningful responses which are a valid reflection of their views.
Forward-focussed	Although methodological options should be founded on a review of best practice to ensure that existing methodologies are well understood, the focus of the study should be forward-looking, and consider creative new research ideas to inform development of ODI rates alongside these.
Robust	Results generated by the recommended approach should be valid and reliable, i.e., unbiased and not highly variable with respect to repeated measurement.
Comparable	Whilst there may be a need for differences across companies, for legitimate reasons such as different languages or different legislative contexts, the methodology should aim to produce comparable results across companies.

A number of additional aspects were raised during the consultation as being important. These included that the methodology should also be:

Inclusive	The sampling methodology adopted should be inclusive: capturing hard-to-reach, vulnerable, and future customers.
Flexible	The ODI research methodology should have the flexibility to accommodate ongoing PC developments
Proportional	Sample sizes and efforts need to be proportionate to the materiality of the metrics.
Timely	The methodology should ensure timely delivery so that companies can plan for applications of the results in their business plan decisions
Deliverable	Any decisions reached regarding survey mode and sample sizes would be deliverable by small companies.
Smartphone compatible	Survey exercises should be designed to work well on a smartphone since this is how the majority of online surveys are now completed.

These principles have been central to the development of the proposals put forward in this report.

3 Survey design

This section presents recommendations for the survey design and the strengths and weaknesses of the options considered in developing these recommendations. It includes the following parts:

- Background and objectives (3.1)
- Stated preference design options (3.2)
- Recommended state preference approach (3.3)
- Selection and definition of service issues (3.4)
- Choice formats (3.5)
- Experimental design (3.6)
- Issues concerning the value measure obtained (3.7)
- Questionnaire structure. (3.8)

3.1 Background and objectives

The core requirement of the Collaborative ODI research is that it obtains measures of how variations in service levels are valued for the purposes of setting ODI rates. In developing the survey design recommendations, we have accordingly drawn from the academic and key policy guidance literature on valuation approaches, the valuation research conducted by water companies for PR19, and by energy companies for RIIO-2, as well as recent relevant Future Ideas Lab submissions.

Additional significant reference points have included:

- The UKWIR 2011 'Carrying out willingness to pay research' study, (NERA-Accent, 2011).
- Ofwat's expectations for customer research (Ofwat, 2021, Appendix, Annex 1).
- The CCW-commissioned study: Blue Marble (2020) 'Engaging water customers for better consumer and business outcomes'.

Notwithstanding the importance of these reference points, the survey design development has been forward-focussed, in line with one of the core principles for success set for the methodology. Accordingly, we have taken on board the steer to consider creative new research ideas to inform development of ODI rates rather than limiting attention to existing methods.

3.2 Stated preference design options

This section reviews the core stated preference methodologies capable of obtaining the valuation measures necessary to support the setting of ODI rates, including their strengths and weaknesses. These methods include:

- Discrete choice experiments plus contingent valuation
- Best-Worst Scaling/MaxDiff
- Compensation-based valuation
- Menu/Slider choices.

Option 1 – Discrete choice experiments plus contingent valuation

Since SP methods first began to be widely used for water and wastewater services valuation, the most widely used approach has been to combine a small number of 'lower level' discrete choice experiments, each containing 3-6 attributes, with a contingent valuation or package choice exercise to calibrate the values obtained to a broad-ranging realistic package of service improvement. This method was used by Yorkshire Water at PR04 (Willis et al. 2005), and by all water companies subsequently, including almost all at PR19. The approach was also recommended by the UKWIR 2011 guidelines on carrying out WTP research. Accordingly, it is the first option we have considered for the collaborative ODI research.

Figure 3 shows an example of a lower level exercise from PR19, while Figure 4 shows an example package contingent valuation question. Participants would answer a series of between four and eight choice questions in each lower level exercise followed by one or two package valuation questions.

In the lower level exercises, the levels taken by each of the attributes in each of the options vary at each choice question (except where a 'Current service level' option is shown, which would stay fixed for the duration of the exercise). The combinations of levels are determined by an experimental design focused on ensuring unbiased and statistically efficient choice data.

The dataset obtained from a sample of customers completing a choice exercise such as this allows an econometric model to be estimated that quantifies the relative value of each of the attribute levels. If one of the attributes is the bill impact, as is the case in Figure 3, then a monetary measure of WTP can be derived for each attribute level, in relation to a base service level.

Figure 3: Example of lower level discrete choice experiment question

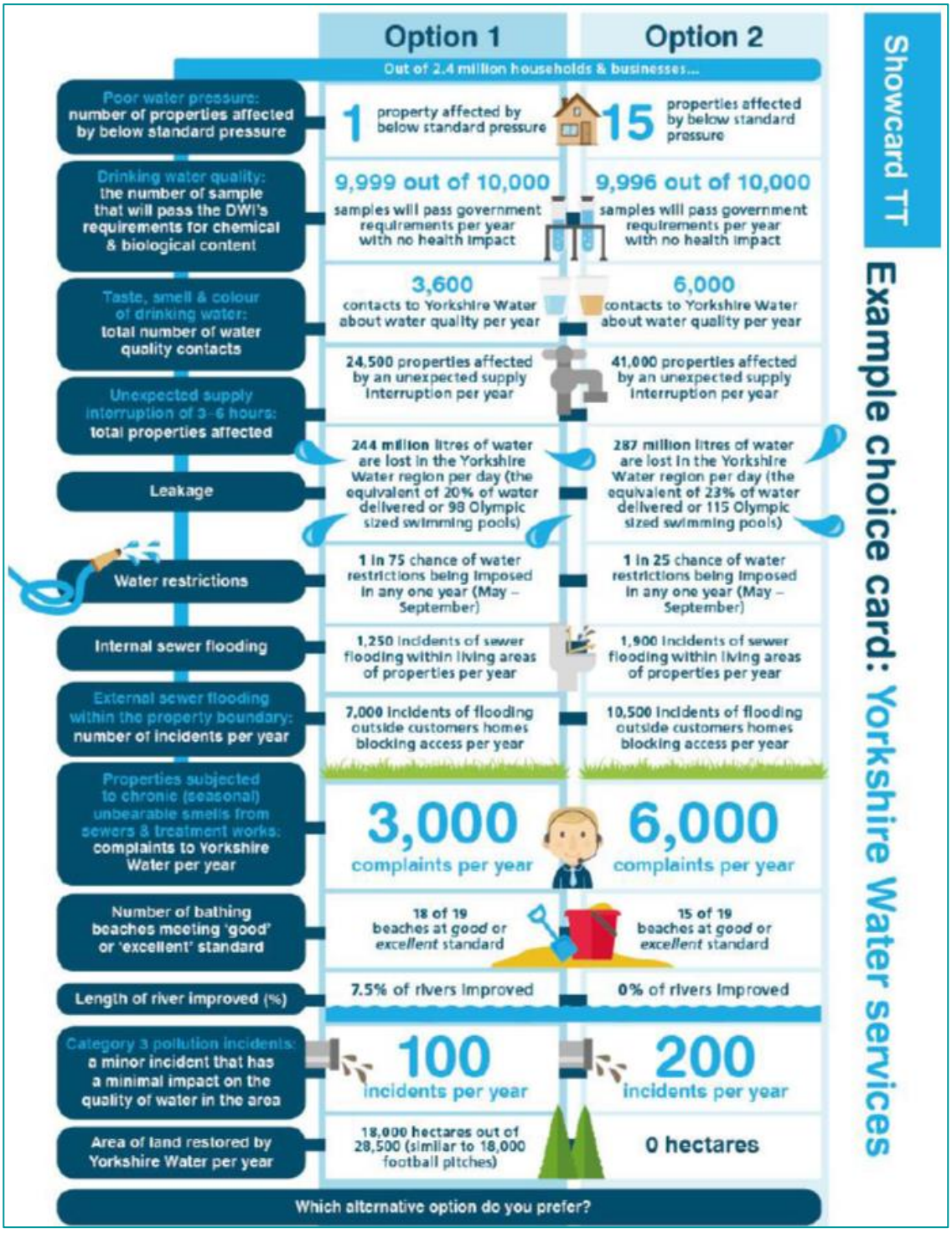
Source: Aecom-DJS (2018) for Yorkshire Water

Typically, and as recommended by the UKWIR 2011 guidelines, water company SP surveys have included a package valuation exercise in addition to the lower level exercises even where bill impacts have been included as attributes of the lower level questions. This is due to the so-called 'package effect', whereby independent valuation and summation of attribute values leads to higher estimates for the value of a package of service improvements than where the package is valued in its entirety.

A package contingent valuation exercise is therefore included as a means of calibrating the scale of the monetary valuations of individual attributes to be consistent with the value of a realistic, but stretching, package of serve enhancements. The lower level choice exercise analysis thus determines only the relative valuations under this approach, while the overall scale of values is determined via analysis of the package exercise choice data.

Figure 4 shows an example package choice question from a PR19 valuation survey. Here, Option 1 is better on every attribute than Option 2, but participants are told that this option costs more, with the costs varying across the sample in order to trace out the willingness to pay distribution.

Figure 4: Example of package contingent valuation question



Source: Aecom-DJS (2018) for Yorkshire Water

Strengths and weaknesses

Table 3 outlines the strengths and weaknesses of using choice experiments combined with package valuation, in comparison to the other stated preference methods reviewed.

Firstly, a strength of the method is that it is well established for valuation of water and wastewater services, having been used since PR04, and throughout PR09, PR14 and PR19. There are also many applications of the method in other sectors, and countries.

The format is also well-suited to allow PCs to be included as attributes without any need for transformation, unless such transformation is deemed desirable for the purposes of ensuring a more consumer-friendly presentation, which is also possible. Other methods require PCs to be transformed to fit within a particular format, which can introduce error.

Finally, the format allows any choice of levels, including enhancements and deteriorations. This thereby allows measurement of both WTP and WTA.

Table 3: Choice experiments: strengths and weaknesses

Strengths	Weaknesses
■ Well established technique for water and wastewater services valuation ■ Allows measurement of PCs in their existing form without the need for transformation/mapping. ■ Allows measurement of both WTP and WTA	■ Choices are complex for participants ■ Choices may be insufficiently sensitive to the range of values shown ■ Choices are too detailed to work well on a mobile phone. ■ The need for package scaling causes difficulties when adding additional measures from other sources for enhancement cases and/or bespoke ODIs.

Note: The number of bullet points on each side does not necessarily indicate the relative strengths of the method

Set against these strengths are some key weaknesses. Firstly, the scenarios generally appear complex to participants, which can result in choices that do not validly measure true preferences. For example, some adopt simple heuristics such as focusing on only one attribute or considering only which levels are better or worse rather than how much each level really means to them.

Related to this issue, a critical issue with respect to this method is that responses are insufficiently sensitive to the range of levels shown. The result is that the resulting valuations are themselves highly sensitive to the service levels included in the design. This is often referred to as the 'denominator effect' (see Box 1).

As has been shown by Metcalfe and Sen (2021), sensitivity to the scope of service change offered in a survey was responsible for the majority of the variation in WTP values observed in the water sector at PR14. This sensitivity to scope is a critical weakness of the method as it implies that the results are unreliable with respect to alternative reasonable survey designs and hence do not represent a valid measure of value. Additionally, it results in excessive valuations where service changes are small, implausible relative valuations across service measures, and implausible relative valuations between companies for the same service measures.

A third weakness of the method is that the choice scenarios are too detailed to work well on a mobile phone. This issue was less significant at the time the UKWIR 2011 guidelines were developed as smart phone prevalence was much lower than it is currently. However, the majority of online survey participants now complete surveys on their mobile phones and so surveys that do not appear well on a mobile phone are likely to put people off from answering, thereby limiting the sample included in the research.

Accordingly, the appearance of the choice questions on a phone is now an important design factor to consider for current and future valuation surveys.

A final issue arises from the fact that the approach requires the use of the package exercise to calibrate values to be consistent with a stretching package of service enhancements. Given that the Collaborative ODI research is focused on common PCs only, companies may need to derive values separately for bespoke PCs. If these values are simply added on to the values derived for common PCs, this could potentially lead to an overvaluation due to the fact that the estimated value for the full package could exceed customers' overall WTP. This is because it is generally the case that adding the values obtained from two surveys, e.g. one covering bespoke PCs and one focused on common PCs, will exceed the value for the combined package when obtained from a single survey, due to less-than-proportional scope sensitivity in willingness to pay valuations as derived from choice experiments and contingent valuation.

Box 1: Scope insensitivity and the 'Denominator effect'

WTP evidence obtained from choice experiment methods that follow the UKWIR 2011 guidance has been found to be highly sensitive to the scope of service change offered in the surveys used to generate the evidence (Metcalfe and Sen, 2021; United Utilities, 2021).

Essentially, the value per avoided service failure is determined by measuring the value for some reduction in the risk of that service failure, say a reduction from 0.5% to 0.4%, and then dividing the resulting value through by the size of the risk reduction, i.e. 0.1%. Thus, if the value measured for the risk reduction is £1 per household per year, the total household value per avoided service failure would be calculated as £1/0.001 = £1,000.

The problem is that the numerator, £1, does not vary in customer surveys in line with the denominator. If one doubles the size of the risk reduction shown to 0.2%, the customer valuation does not tend to double but, rather, it increases only very marginally. In this example, if it stays the same at £1 per household per year, the total household value per avoided service would halve to £1/0.002 = £500.

The denominator effect implies that the results are unreliable with respect to alternative reasonable survey designs and hence do not represent a valid measure of value. Additionally, it results in:

- Excessive valuations where service changes are very small.
- Implausible relative valuations across service measures
- Implausible relative valuations between companies for the same service measures.

The discrete choice experiment plus contingent valuation approach merits attention due to its prominence in past research, and in the broader literature on non-market valuation. However, the approach overall has critical weaknesses in the form it has been applied in the water sector, including that it has led to the lack of comparability that has been observed in successive price reviews, and which was a primary motivation for conducting the Collaborative ODI research to begin with.

Option 2 – Best-worst scaling/MaxDiff approaches

At PR19, several companies implemented an alternative stated preference approach based on the best-worst scaling / MaxDiff methodology[1]. This appeared in two broad forms:

- Service change trade-offs
- Relative impact trade-offs

Figure 5 shows an example best-worst question of the first kind. Here, rather than choosing a preferred combination of service levels as in a discrete choice experiment, the participant chooses which service change they consider the best for them, and which is the worst.

Figure 5: Example MaxDiff question (Service change trade-off version)

	CURRENT SITUATION	SCENARIO 1	BEST (CHANGE IN SERVICE LEVEL)	WORST (CHANGE IN SERVICE LEVEL)
Unplanned interruptions *Number of properties affected by **unplanned interruption to water supply** (6-12 hours) each year*	18,000 properties (1 in 100 properties)	BETTER 14,000 properties (1 in 120 properties)	○	○
Severe water restrictions (Rota cuts) *How often **severe water restrictions could be experienced***	1 in 100 years (25% chance in the next 25 years)	WORSE 1 in 25 years (100% chance in the next 25 years)	○	○
Discolouration *Number of properties affected by **discolouration of tap water** each year*	30,000 properties (1 in 60 properties)	BETTER 10,000 properties (1 in 170 properties)	○	○
Leakage *Percentage of **water lost due to leakage** each year*	15% (1 in 7 litres)	WORSE 22% (1 in 5 litres)	○	○

Source: ICS-Eftec (2018) for Anglian Water.

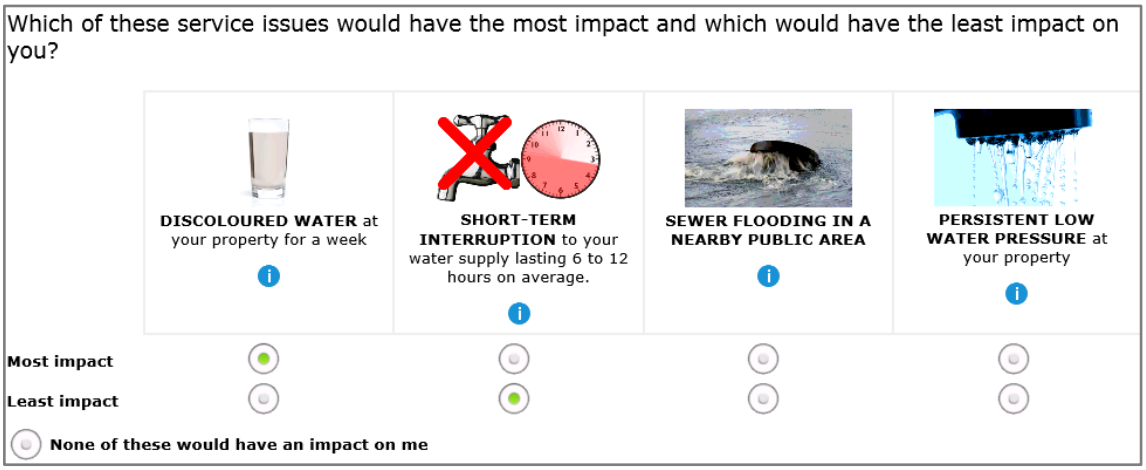
Answering a sequence of questions like the above returns a dataset that allows for the estimation of a quantitative index of relative utility. To convert this into a WTP measure, an additional package choice contingent valuation exercise is needed that includes a trade-off between bill levels and service improvements (e.g. Figure 4).

Separately, the impact-trade off version of the MaxDiff approach focuses on deriving an index of relative impact over all the service issues rather than between service level changes. An example of this kind of question is shown in Figure 6. The key difference here is that the participant is asked to focus on the service issues only, and not to consider service levels per se.

In this exercise, participants are shown a sequence of questions where the four service issues varies each time. To convert this into a WTP measure, an additional exercise is again needed that includes a trade-off between bill levels and service improvements (see e.g. Figure 4). In this case, however, the mapping from package WTP to individual service-level WTP involves imposing the restriction that values for service level enhancements are proportional to the impact-weighted number of customers impacted by the service level change. (See Chalak and Metcalfe, 2021, for details.)

[1] See Louviere et al (2015) for a textbook treatment of this methodology.

Figure 6: Example MaxDiff question (Relative impact trade-off version)



Source: Accent-PJM Economics (2017) for Dŵr Cymru Welsh Water.

Strengths and weaknesses

Table 4 summarises the strengths and weaknesses of the above methods. Best-worst scaling/MaxDiff methods have the key advantage over discrete choice experiments of generally being simpler for participants to answer. This is because they don't have to evaluate whole packages of attributes at a time. This advantage is particularly emphasised in the impact-based version which involves no trading off of service levels at all.

In general, the method is also more statistically efficient than discrete choice experiments, which results in more precise estimates for the same sample size (Louviere et al, 2015). These advantages are gained within the same theoretical framework (Random Utility Theory) as the more traditional discrete choice experiment approach.

When considering the impact-based approach in particular, the method has additional strengths including, primarily, that it avoids the need for participants to have to trade off small risk changes against one another. This feature is primarily responsible for the cognitive complexity of the discrete choice experiment approach when applied to water and wastewater services, and is the cause of the denominator effect (see Box 1) with all its concomitant issues.

It does so by imposing the principle of rational choice under risk, also known as 'Expected Utility Theory' (Von-Neumann and Morgenstern, 1953). Under this principle, the utility, or value, that a customer obtains from a small reduction in the risk of a service failure is equal to the change in probability multiplied by the loss of utility associated with the service failure itself. For example, a reduction from "2 in 10" to "1 in 10" in the chance of a supply interruption would be valued as (2/10 -1/10) times the value of an avoided supply interruption. This makes the valuation potentially much simpler because one only needs to know the value of an avoided supply interruption to value any reduction in the risk of a supply interruption.

The advantage of imposing this principle in the case of water company WTP surveys is that one can focus on the relative aversion to different types of service failure directly,

and then simply scale these values by the extent of the service improvement, measured by the change in risk per customer, in order to measure the relative value of each of the candidate service improvements. In other words, there is no need to ask respondents to consider small risk changes in order to understand the relative values of different types of service improvement.

Since service levels are taken out of the choice question in the impact-based approach, the questions themselves will typically take up less screen space and can accordingly work well on a mobile phone. This represents an important additional advantage to the method.

Finally, the impact-based approach can accommodate more service issues than a discrete choice experiment within a single survey. For example, multiple different durations and severities of service incident can be explored and their relative impacts compared. By contrast, when using the traditional discrete choice experiment approach, the relative impacts of different severities have often been explored via a second 'Stage 2' survey. Including them within a single survey can therefore provide a less costly method of obtaining the same granularity of valuations.

Table 4: Best-worst scaling / MaxDiff: strengths and weaknesses

Strengths	Weaknesses
■ Simpler for participants to answer, especially the impact-based version. ■ Statistically efficient in comparison to discrete choice experiments. ■ Based on the same underlying theory (Random Utility Theory) as choice experiments, and with equal academic support. ■ With the impact version: – Avoids the need for participants to deal with small risk changes. – Choices can be set up to work well on a mobile phone. – A greater number of service issues can be included.	■ Not possible to measure interaction effects between PCs. ■ Requires a separate exercise to convert utilities to monetary values. ■ With the service change trade-off version: – Choices may be insufficiently sensitive to the range of values shown. – Choices are too detailed to work well on a mobile phone. ■ With the impact version: – All PCs must be translated to a form amenable to evaluation of relative impact. – Imposes a 'rational' structure to valuing risky prospects, which may not be consistent with what people would freely choose for themselves.

Note: The number of bullet points on each side does not necessarily indicate the relative strengths of the method

The weaknesses of the method again depend on which version is considered. A core limitation of both versions, however, is that they are unable to measure interaction effects between attributes. For example, in the impact-based version, there is no possibility of measuring the joint impact of experiencing both a short-term supply interruption and sewer flooding in a nearby public area, to test whether this is greater or less than the sum of the two impacts individually. This is unlikely to be a significant limitation in the case of water and wastewater services valuation research. Even when

using the discrete choice experiment method, which does allow the estimation of interaction effects, it is extremely rare for studies to model these effects and take them into account when making recommendations for appraisal values. In the vast majority of cases, only the main effects of attributes are modelled and valued.

A more significant general feature of the methodology, which could potentially be a limitation, is that it requires a separate exercise to convert utilities into monetary values. At PR19, this involved using a package contingent valuation exercise, of the kind shown in Figure 4. (See Chalak and Metcalfe, 2021, for details of how this is combined with the impact-based exercise to derive monetary values for individual service issues.) Given the amount of detail included in such an exercise, some of the benefits of the best-worst scaling/MaxDiff approach are foregone, including the possibility of presenting the material on a mobile phone.

Moreover, there are likely to be scope sensitivity issues affecting valuations from the package contingent valuation exercise, due to its complexity, in the same way as there are from the discrete choice experiment approach more generally. This means that the overall size of the valuations is still likely to be dependent on the scope of service change offered overall. Although, this should be selected to be consistent with a realistic package of improvement, there could be different sizes of improvement package that are considered realistic and so WTP values may still not be fully reliable with respect to alternative reasonable designs at the package level.

The remaining weaknesses of the method are dependent on which version is being considered. With the service change trade-off version, the method still requires participants to trade off small changes in service levels against one another, which can be expected to lead to choices that are insufficiently sensitive to the range of values shown. This is for the same reason that discrete choice experiment results are similarly affected (see Box 1). Additionally, choices are too detailed to work well on a mobile phone with this version.

With the impact version, a key limitation is that all PCs must be translated to a form amenable to evaluation of relative impact. It is impossible to include a PC such as leakage reduction, for example, as a continuous measure. Instead, the method requires that there must be a direct impact of some kind, that customer either experience or they do not. In the case of leakage, this requires consideration of how leakage actually does impact the customer, which can be debated. On the other hand, by enforcing the discipline of identifying how a PC impacts on customers, and switching the focus to measuring that impact, it is arguable that the resulting valuation is more meaningful as a consequence.

The final potential limitation of the impact-based method is more philosophical in nature in that it imposes a rational structure to valuations which is less likely than other approaches to be consistent with what customers would freely choose for themselves. The approach is therefore arguably less consistent with the principle of consumer sovereignty. Whether this is considered to be a concern depends on one's perspective on the role of appraisal – is its purpose to achieve an outcome consumers would choose for themselves, even if that choice is fraught with difficulties, or is its purpose to maximise welfare as experienced by the consumer? Cost-benefit analysis adopts both perspectives

at different times, although most often little attention is given practically to such philosophical questions [Smith and Moore, 2010].

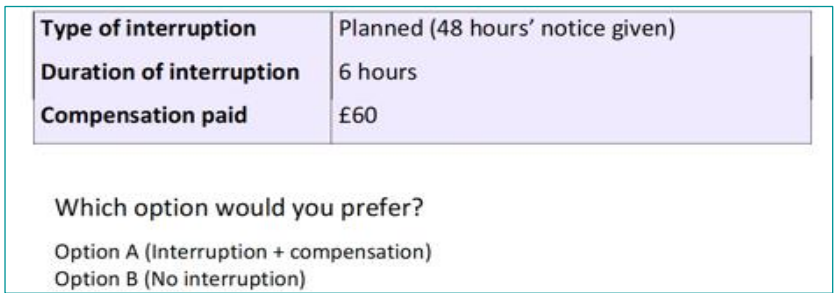
Option 3 – Compensation-based approach

At PR19, Affinity Water commissioned a study focused on the valuation of supply interruptions via exploring the levels of compensation that customers would like to see given to those experiencing supply interruptions. Required compensation amounts are, in principle, a valid measure of value for cost-benefit analysis and, accordingly, also for ODI rate setting.

The key question addressed by this research was which level of compensation payment was needed to fully compensate customers for the inconvenience of a supply interruption. This is the amount above which customers would, on average, prefer to deal with the inconvenience of the interruption and receive the compensation payment rather than not have the interruption in the first place.

Figure 7 shows an example of the type of question asked in the Affinity Water study. In that study, a sequence of questions like this was asked of each participant, with the type and duration of interruption, and the compensation paid, varying each time according to an experimental design. This resulted in a dataset of choices that could be used within an econometric analysis to determine mean and median willingness to accept compensation values per hour of interruption avoided, for both planned and unplanned interruptions.

Figure 7: Example compensation-based choice question



Source: Accent-PJM (2018a) for Affinity Water.

The compensation-based approach could play one of two potential roles:

- It could serve as a multiple attribute valuation exercise on its own, like the discrete choice experiment approach, in which case many different types of service issue would need to be included, but no additional exercise would be needed.
- It could be used alongside an impact-based approach (e.g., Figure 6) as a means of monetising the impact index estimated via that approach. To do so, 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.

For example, if we know from the compensation-based approach that the value of an avoided 6-hour supply interruption is £50 and, from the impact-based MaxDiff exercise, that having discoloured water for two days is three times as impactful, then the derived value for an avoided discoloured water incident lasting two days is £150.

Strengths and weaknesses

Table 5 presents a summary of the strengths and weaknesses of the compensation-based approach. In comparison to the traditional discrete choice experiment approach, the method offers a number of advantages whilst still remaining consistent with economic valuation theory.

Firstly, the questions are simpler and less abstract for participants. This is clear from simple inspection of Figure 7,and is driven by the fact that participants are not required to consider packages of service levels, nor even to consider service levels at all. This has the critical benefit that the denominator effect is avoided altogether in this approach (see Box 1). It also has the procedural benefit that service levels will not need to be agreed for every company in a short space of time as part of the Collaborative research design.

A further advantage of a simpler format is that the format can work well on a mobile phone.

Table 5: Compensation-based approach: strengths and weaknesses

Strengths	Weaknesses
■ Theoretically valid measure of the value per avoided service issue. ■ Simpler and less abstract for participants to answer. ■ Avoids the need for participants to deal with small risk changes, and hence avoids the denominator effect (see Box 1). ■ Avoids the need for service levels at all, simplifying the process of survey design for the Collaborative research. ■ Choices can be set up to work well on a mobile phone.	■ Only tested in the water sector for supply interruptions. Accordingly, unknown if it would work for all the different types of service issues that would need to be valued. ■ All PCs must be translated to a form amenable to compensation questioning. This may be problematic for many PCs. ■ Method relies on levels of compensation shown being perceived as credible, which may be problematic where current compensation levels are well below required levels. ■ Method does not set any bounds on overall willingness to pay for bill increases.

Note: The number of bullet points on each side does not necessarily indicate the relative strengths of the method

The weaknesses of the method are dependent, to a degree, on which of the two potential roles the exercise is intended for in the PR24 Collaborative ODI research. In the water sector the method has only been used to date to value avoided supply interruptions. It is accordingly untested for all other types of service issue that would need to be included in order for the approach to be usable on its own to set ODI rates.

There are important reasons why the approach may fail to work for all the required PCs. Even if the PC can be mapped to a service issue, as is also required by the impact-based MaxDiff approach, the service issue may be incompatible with a compensation-based approach. This could be either due to the fact that compensation is inappropriate or not credible as a private payment for certain service issues, such as those that impact the environment; or, that the amounts of money needed to fully compensate people for some service issues, such as internal sewer flooding, may be so large as to be considered incredible by participants, particularly given the fact that real-world compensation payments are so low by comparison.[2]

A final possible limitation of the method, in comparison to methods that include a package exercise, is that it does not directly give evidence on how much customers are willing to pay for a package of improved service levels. Without any limitation on this, the ODI rates thus generated could lead to a situation where a company raises bills, in line with service improvements, beyond the level that most customers would be prepared to pay. In order to avoid this situation, it could therefore be appropriate to undertake an additional, separate, piece of research including an exploration of customer preferences in this area. (This issue, and the type of research envisaged, are discussed further later in this report.)

Overall, the compensation-based approach appears to have much to commend it, particularly as a complement to an impact-based approach (e.g. Figure 6) in place of the package contingent valuation exercise (e.g. Figure 4) which has been used in previous studies using this approach.

Option 4 – Menu/Slider choice exercise

The fourth and final stated preference approach considered for obtaining valuations is the menu-based choice exercise. In this approach, survey respondents construct their own package of attribute levels from a menu where each level is associated with a certain cost.

Many companies utilised this approach at PR19. In some forms of presentation (e.g. Figure 8), service level choices appear as discrete levels, whereas in other forms (e.g. Figure 9), sliders are shown which allow participants to choose on a continuous scale where they would like to set service levels. In both versions, the bill impact of the participant's choices is updated in real time so that the participant can adjust all levels to be consistent with the bill increase they would be willing to pay.

[2] Under the guaranteed standards scheme (GSS), customers that experience an internal sewer flooding incident are entitled to compensation equal to the size of their annual sewerage bill by wastewater companies. The average sewerage bill in 2021/22 is £213 (discoverwater.co.uk/annualbill). By contrast, average household WTP value per avoided internal sewer flooding incident at PR19 ranged from £1,772 to £123,477 (Accent-PJM Economics, 2018b).

Figure 8: Example of menu choice exercise

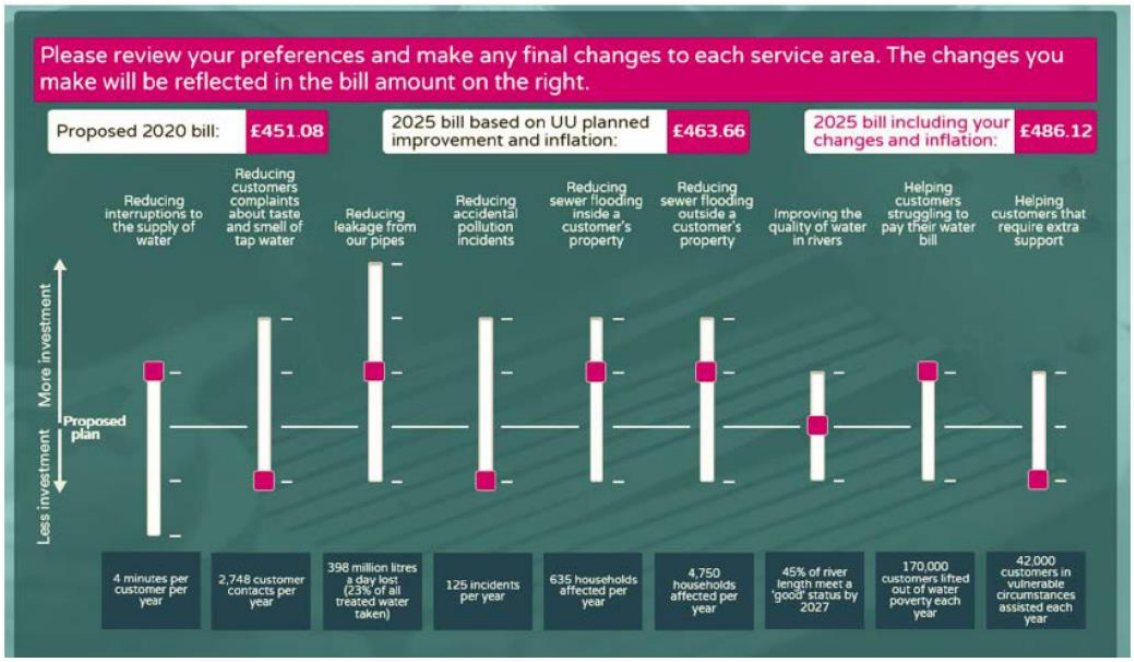
	Option A	Option B	Option C	None	Cost
Drinking water acceptability *(Contacts per 1,000 population)*	☑ 2.3	☐ 2	☐ 1.6	☐	No change to your bill
Drinking water availability *(Average minutes lost)*	☑ 12.2	☐ 10	☐ 7	☐	No change to your bill
Leakage *(Litres/property/day)*	☐ 121	☑ 117	☐ 114	☐	+£0.66 every year for 5 years
Preventing pollution *(Number of cat 3 incidents)*	☐ 103	☑ 90	☐ 70	☐	+£2.5 every year for 5 years
River water improvements *(km improved)*	☐ 0	☑ 150	☐ 225	☐	+£2.5 every year for 5 years
Sewage in the home *(Properties)*	☐ 225	☐ 200	☑ 180	☐	+£0.45 every year for 5 years
Sewage in the street *(Properties)*	☐ 6500	☑ 6300	☐ 6100	☐	+£1 every year for 5 years
Worst served customers - low pressure *(Properties)*	☑ 35	☐ 10	☐ 0	☐	No change to your bill
Worst served customers - interruptions to supply *(Properties)*	☑ 1400	☐ 1000	☐ 800	☐	No change to your bill
Worst served customers - sewer flooding *(Properties)*	☑ 1648	☐ 1250	☐ 1000	☐	No change to your bill
Help for disadvantaged customers *(No. customers on social tariffs)*	☐ 100,000	☐ 150,000	☑ 200,000	☐	+£0.33 every year for 5 years
Resilience of wastewater networks to storms *(Roof equivalents)*	☑ 25000	☐ 40000	☐ 60000	☐	No change to your bill
Reducing fossil fuel dependency *(% of total energy use)*	☐ 30%	☐ 35%	☑ 40%	☐	+£2.5 every year for 5 years
Protecting your service in extreme events *(% Resilience)*	☐ 84%	☑ 87%	☐ 90%	☐	+£1.25 every year for 5 years

Water Bill

Your choices result in total bill change of **+£11.19** (This would take your bill from £520 to £531.19)

Source: Accent (2017) for Dŵr Cymru Welsh Water.

Figure 9: Example of slider choice exercise



Source: BoxClever (2018), for United Utilities.

At PR19, the bill impacts of varying service levels tended to be fixed at their true expected rates, based on marginal costs, rather than varied across the sample. This means that the information gained from the sample as a whole is limited to the proportion that are willing to pay the estimated cost for each service level. Whilst this information is potentially useful, it does not in itself provide reliable evidence on the overall value of service increments and decrements. It is hence unsuitable as a method for the Collaborative research.

For the menu/slider choice exercise to be suitable for obtaining the reliable value evidence needed to inform ODI rates, it is necessary for there to be experimental variation in the marginal costs shown across the sample. This variation could be used, in principle, to trace out how demand varies with cost, and thereby reveal valuations for service changes via an econometric analysis. However, this approach is untested. At PR19, only one company (Yorkshire Water) reported having varied the cost levels across the sample and, in this case, the report on the study did not contain any estimates of willingness to pay (Aecom, 2017).

Strengths and weaknesses

Table 6 contains a summary of the key strengths and weaknesses of menu-based choice experiments. In the context of water and wastewater services, the key strength of the method is that it gives participants the ability to choose their preferred service levels, which is something that many participants like to do. The survey experience can thereby be a more positive one for participants than traditional choice experiments or other methods that require them to choose between options when they might not like any of them.

However, to our knowledge, the approach has not been used for non-market valuation within the academic literature. Furthermore, where companies derived measures of WTP from studies at PR19 based on designs with no marginal cost variation, the resulting measures cannot be considered reliable estimates of marginal benefits. Whilst possible, in principle, an approach involving experimental variation in costs combined with an econometric analysis to estimate the demand curves for each service area, is untested and, arguably, carries a risk that it could fail to recover reliable measures of WTP.

Table 6: Menu/Slider approach: strengths and weaknesses

Strengths	Weaknesses
■ Allows participants to choose their own preferred service package, which they often like to do.	■ When used without experimental cost variation, evidence on WTP is unreliable. ■ Estimation of WTP via an approach using experimental cost variation is untested, and carries a risk that it could fail to recover reliable measures of WTP. ■ Still requires trading off of small changes in service levels. ■ Format does not work well on a mobile phone.

Note: The number of bullet points on each side does not necessarily indicate the relative strengths of the method

A further key weakness of the method is that it still requires participants to trade off service levels against one another, and so is liable to be sensitive to the denominator effect (see Box 1) that also afflicts traditional choice experiment methods.

Finally, the method requires a lot of screen space and is hence unsuitable for use on a mobile phone.

Overall, therefore, the menu/slider approach appears to be unsuitable as a method worth pursuing for the Collaborative ODI research. This conclusion could, in principle, be overturned for future research if a version of the approach is developed that can demonstrate superior validity and reliability than the recommended method. However, in our view, this is unlikely to be the case due to its continued reliance on the need for participants to trade off service levels against one another.

3.3 Recommended stated preference approach

Based on the strengths and weaknesses of the selection of options reviewed in Section 3.2, our recommendation for the Collaborative ODI research is that the core survey should be designed around two choice exercises:

- A compensation-based choice exercise to estimate the values of two different types of avoided service failure.
- An impact-based exercise to estimate the relative impact of all types of service failure.

These results would be combined in the analysis to return estimates of the values of all types of service failure. For example, if we know from the compensation-based approach that the value of an avoided 6-hour supply interruption is £50 and, from the impact-based exercise, that having discoloured water for two days is three times as impactful, then the derived value for an avoided discoloured water incident lasting two days is £150.

We are recommending that the compensation-based choice exercise includes two service issues, and hence two pivot values, rather than a single service issue as originally conceived. This is in order to mitigate against any over-reliance on a single estimate. (This suggestion was made within the 13 Dec 2021 industry workshop when the approach was presented and discussed. See Appendix B for further details of stakeholder views.)

The two values obtained from the compensation-based choice exercise would each serve as pivot points, thereby returning two sets of values in total. This will define a range for each valuation rather than a single fixed point, and will allow for some understanding of the sensitivity of valuations to the choice of service issue used to determine the pivot valuation. (See Section 5 for details of the expected analysis and outcomes, including worked examples.)

The key advantages of this combined approach are:

- It is simple and customer focused, whilst remaining consistent with the requirements for economic valuation evidence needed to support the setting of ODI rates.
- It avoids the need for service levels at all, thereby avoiding the denominator effect altogether and its concomitant issues (see Box 1), plus the practical problems associated with companies needing to provide estimates of base service levels and variations around these within a short space of time.

- If there are changes to PC definitions following completion of the core survey, provided the service issues have been carefully chosen to span the range of impacts that water and wastewater services could have, the results should be flexible enough to provide the valuation evidence needed for the new definitions. (Further details on the selection of service issues to be included in the survey, and how they are intended to be mapped to the common PCs, are discussed in Section 3.4.)

- Choices can be set up to work well on a mobile phone.

- Where additional service issues require valuation for the purposes of bespoke ODIs and / or enhancement cases, the method allows for valuations to be obtained in separate studies without any issues caused by the need for package scaling. (See Section 5.5 for discussion of how company research should be designed to link into this methodology.)

The approach raises a number of issues that require further discussion:

- How to select and define the service issues to be used in both the impact-based and compensation-based choice exercises, and understand any potential limitations?

- How should the choices be presented in the survey?

- How should the experimental design be created?

- Is the right measure of value obtained by this method?

    - Is there a risk of over-valuation due to the use a WTA measure of value in the compensation-based approach, rather than a WTP measure?

    - Are there differences between customer preferences as individuals and as citizens, and how are these taken into account?

These issues are discussed in the remainder of this section, followed by a summary of the recommended full questionnaire structure.

3.4 Selection and definition of service issues

The recommended SP method requires that PCs must be mapped to customer-facing measures in all cases rather than necessarily being used in the same form as set for regulatory purposes. Specifically, it requires that every attribute that is valued in the research should be described as a type of incident that could impact the customer in some way; for example, a supply interruption or an internal sewer flooding incident. This imposes the design challenge to ensure that the marginal value associated with each of the common PCs needed for PR24 can be expressed algebraically as a function of the likelihoods of one or more service issues occurring. Moreover, it requires that these service issues be fully specified to avoid ambiguity as far as possible.

This requirement is substantive. There are some PC measures that do not easily map to customer outcomes; leakage being a prominent example of these. Leakage is a

performance measure that customers do not experience directly, but which often generates an emotionally charged response from the public, and the media. In such cases, and as a general rule, the method proposed here requires that the impacts of varying these PC measures be translated to impacts that will be directly experienced by the public rather than including them directly within the survey.

In support of this general proposal, we note the following with regard to leakage:

- People do not value leakage rates per se. Instead, over myriad customer research studies, including the recent Yonder research for CCW, it has been shown that customers consider leakage to be emblematic of a failing and wasteful water industry, which they perceive as leading to:
  - Higher bills,
  - More water being taken from the environment
  - Unreasonable restrictions imposed on customer water use, either
    - formally through drought restrictions or, more commonly, through
    - requests to use less water themselves.

- If people did value leakage for its own sake, it should be possible to obtain an approximate consistency of valuations across companies, regardless of the levels of leakage reduction included in companies WTP survey designs. As shown in Ofwat (2021a), however, PR19 estimates of the value that customers put on a normalised unit of leakage reduction varied from £0.03 to £41.58, a factor of 1,386. This undermines considerably the idea of an approximately stable underlying value.

Given this, it is appropriate, in our view, for the research to value the impacts of leakage rather than leakage itself. This should obtain a more reliable and meaningful measure of value. Likewise, the general principle that PCs should be translated to customer-focused impacts ensures that they are asked questions that they can answer based on their own direct experience rather than in the abstract, and thereby should also result in more reliable and meaningful measures of value.

The remainder of this section includes the following parts:

- It includes a discussion of the key factors to be included in the definitions of each of the service attributes, in general terms.

- It then addresses each group of common PCs, in turn, focusing on how they may be represented in the survey as a customer-facing service impact and be mapped back from that representation to the original PC definitions. In each case, there remain a series of issues to be addressed in order to finalise the survey and to set the values of the parameters needed to complete the mappings back to the PC definitions.

- Finally, it includes recommendations, with justification, for the selection of the two service issues to be included in the compensation-based choice exercise.

Key factors to be included in service issue definitions

In general, service issues should be specified to include all the key factors relevant to assessing how a customer might perceive its impact. These include:

- Type of issue, and its impact
- Whether any notification is given
- Cause
- Duration, and timing

These factors are considered for each of the service areas in the following.

Customer service PCs

The first category of PCs to be considered in detail contains those based on service experienced at customers' properties. This includes:

- Water supply interruptions
- Drinking water quality, including Compliance Risk Index, Event Risk Index and Customer contacts about water quality
- Sewer flooding, including internal and external types.

Water supply interruptions

Expected PC definition
Average number of minutes lost per customer for the whole customer base for interruptions that lasted three hours or more.

This PC incentivises reducing the number and duration of interruptions over three hours long. It covers both planned and unplanned interruptions.

For valuation, it is only necessary to include one type of interruption in the design to map to this PC. For example, if the survey design valued an unplanned interruption lasting 6 hours, the value per minute against the performance commitment definition could be calculated as N/360 times the value per avoided 6 hour interruption, where N is the total number of customers supplied.

However, the value per minute will not necessarily be the same for all durations, and is likely to be somewhat less for unplanned interruptions than planned interruptions. It is also likely to be higher at certain times of the day than at others. This suggests that some care needs to be taken over the choice of attribute(s) to include in the design.

We have considered two options for the choice of attributes:

1. Include just one 'typical' interruption, corresponding to the modal type and median duration.

2. Include multiple types of interruption and attempt to derive an appropriately weighted average of the resulting valuations.

Within the first option, it would be possible to have the same type of interruption for every company or allow this to vary by company in line with the typical interruption in each company's area. In the interests of comparability, however, we would suggest that it would be preferable to have the same type of interruption shown for all.

With regard to the second of these options, whilst it would require a decision to be made regarding how to weight each of the interruption values, it would provide a richer source of evidence on the value of different types of supply interruption than including a single type only and would accordingly provide more flexibility in the event that the PC definition changes. It would also provide companies with more information that they could use to develop enhancement cases that depend on interruptions values.

Accordingly, in the interests of flexibility, we recommend including four types of interruption:

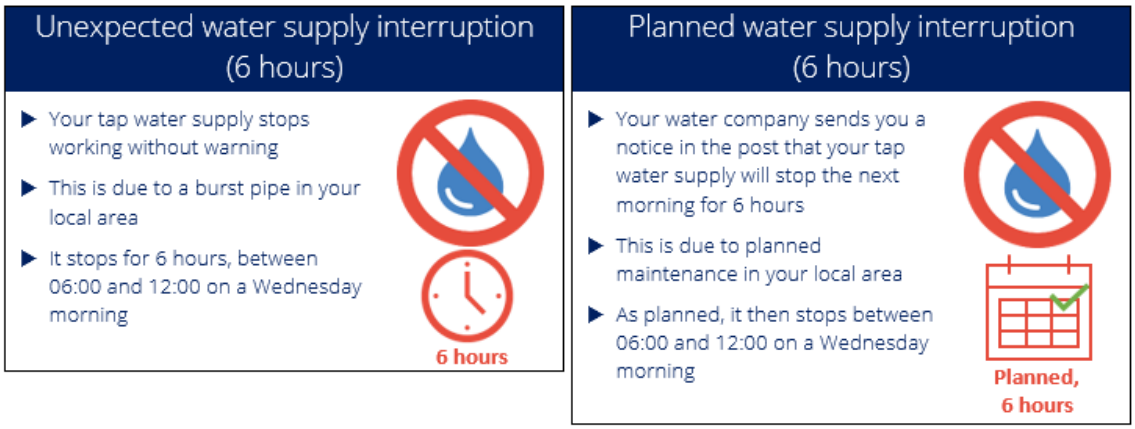
- Unexpected interruption (6 hours)
- Unexpected interruption (24 hours)
- Planned interruption (6 hours)
- Planned interruption (24 hours)

This will provide data on the relative value of planned versus unexpected interruptions, and a measure of how the impact cost to customers increases with the duration of the interruption.

In the definitions of these, it is important to also be specific about the time of day the interruptions would happen, and it can also be beneficial to give a reason for the incident. Rather than expand the number of service issues further to accommodate different times of day, we recommend adopting a single time of day and using this throughout the exercise.

Figure 10 shows a mock-up of how the supply interruptions could look in the survey.

Figure 10: Supply interruptions definitions



Mapping

- 1 minute per customer equals N / 360 6-hr unexpected interruptions, where N is the number of customers in total; or,
- Weighted average of similarly derived values for 6-hr and 24-hr, unexpected and planned interruptions.

Drinking water quality

Expected PC definitions

There could be three common PCs representing drinking water quality, defined by the DWI, (DWI, 2018a, b), including:

- Compliance Risk Index (CRI)
  - The sum of individual compliance failure scores over the course of a year, where each failure is scored by multiplying a Parameter score [1-5] by an Assessment score [0-5] by the proportion of the population (or volume supplied, or reservoir capacity) affected.
- Event Risk Index (ERI)
  - The sum of individual scores for all events notified to the DWI over the course of a year, where each event is scored by multiplying a Seriousness score [0-5] by an Assessment Outcome score [1-5] by an Impact score, calculated as the size of the population affected multiplied by the duration of the incident, in hours, and divided by the population served by the water company.
- Customer contacts about water quality[3]
  - The sum of all communications about drinking water quality initiated by a consumer living or working in the area supplied by the water company including phone, letter, fax, email, in person, website request form and message left on a helpline.

    There are two categories of consumer contact included in this measure, defined as follows:

    - A consumer contact about the appearance of drinking water is a contact where the consumer perceives something different about the appearance of the water from the "norm".
    - A consumer contact about the taste and odour of drinking water is a contact where the consumer perceives that the water has a taste or smell.

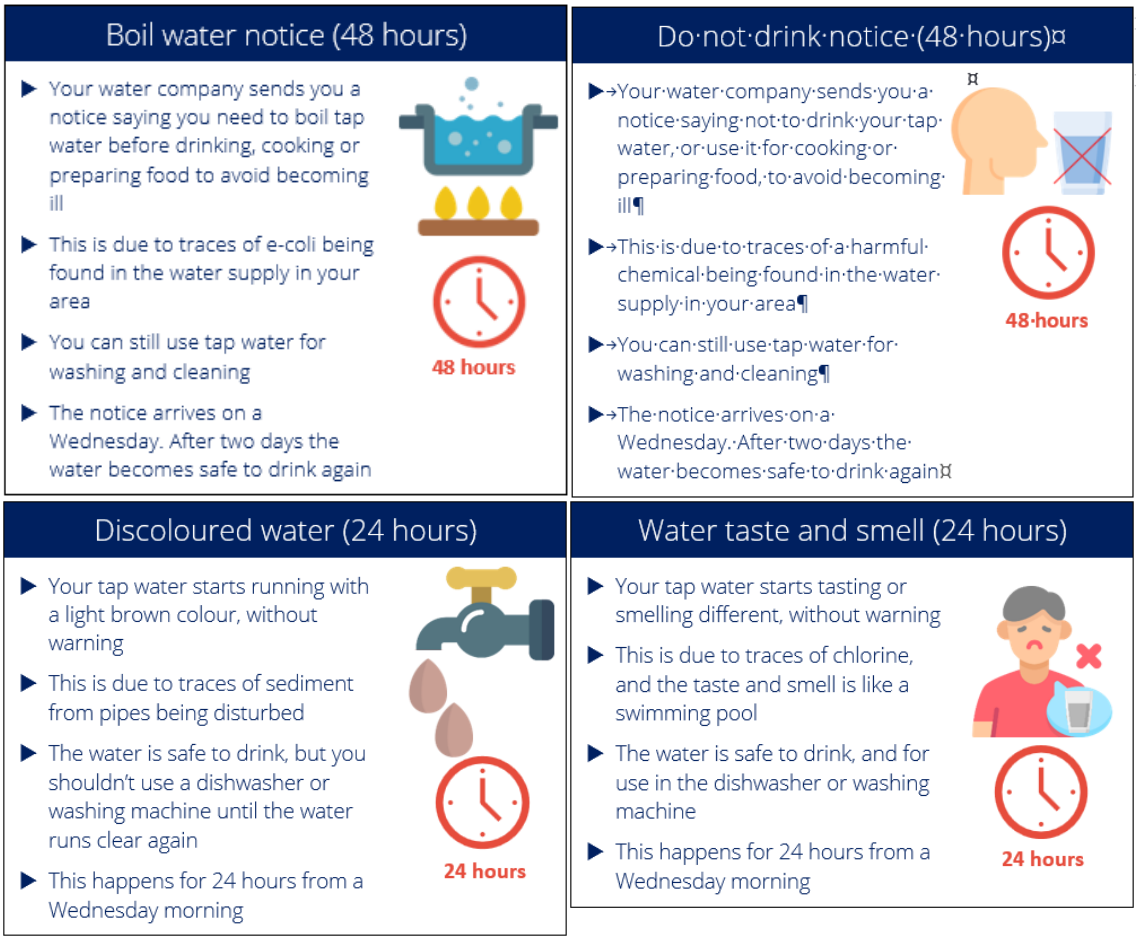
3 DWI (2006) Annual provision of information on consumer contacts. Information Letter 1/2006. 6 Jan 2006.

These measures are overlapping. The CRI and ERI PCs are both intended to incentivise companies to fully comply with statutory obligations and to mitigate any issues affecting the quality of drinking water. The number of contacts about water quality provides a further measure which identifies cases where, although water is safe to drink, it is not acceptable to customers.

Because the measures are overlapping, it is important to ensure that there are no issues due to double-counting. Our proposed general approach to this problem is set out in Appendix A. It shows that in order to value overlapping PC measures, it is necessary and sufficient to agree the impacts of a 1-unit change in each PC, on the likelihoods of each underlying customer-facing water quality issue.

For the case of drinking water quality, we propose that each of the three PCs should be expressed, as best as possible, as a function of the following proposed service issues that the customer could experience. (See Figure 11.)

Figure 11: Drinking water quality issues definitions



NB: In the case of Discoloured water and Water taste and smell, there could be value in including both notified, and unexpected, versions of these service issues. (See below, and Appendix A, for a discussion of this issue)

Mapping

- CRI $= w_{11}r_{BW} + w_{12}r_{DND} + w_{13}r_{DW} + w_{14}r_{TS}$
- ERI $= w_{21}r_{BW} + w_{22}r_{DND} + w_{23}r_{DW} + w_{24}r_{TS}$
- Contacts $= w_{31}r_{BW} + w_{32}r_{DND} + w_{33}r_{DW} + w_{34}r_{TS}$

Where:

r_{BW} is the average annual risk of a customer receiving a boil water notice at their property

r_{DND} is the average annual risk of a customer receiving a do not drink notice at their property

r_{DW} is the average annual risk of a customer experiencing discoloured water at their property

r_{TS} is the average annual risk of a customer experiencing tap water with an unusual taste and smell at their property

And:

$w_{ij,}$ are the impacts on the company's CRI, ERI and Contacts scores per one customer property experiencing the corresponding service issue.

NB: This could be straightforwardly extended to include both notified, and unexpected, versions of Discoloured water and Water taste and smell.

Under this construction, a 1-unit change in the CRI, for example, has a $1/w_{11}$ impact on the average risk of a customer receiving a boil water notice at their property, a $1/w_{12}$ impact on the average risk of a customer receiving a do not drink water notice at their property, etc.

Application of the approach will require expert input to set the values of all of the w parameters.

The weights w_{33} and w_{34} have natural interpretations based on the likelihoods of customers contacting the company upon experiencing discoloured water and unpleasant taste and smell respectively. For example, if 1 in 4 customers who experience discoloured water go on to contact their water company about it, then w_{33}=0.25 times the number of properties served by the company.

A potential issue has been raised to us by one company regarding the appropriateness of mapping customer contacts to the experience of the corresponding service issues. The company in question contended that the number of water quality contacts could be reduced by proactively contacting customers who are at risk of experiencing a water quality issue in the near future. This, it is argued, breaks the link between the PC measure (water quality contacts) and the risk of customers experiencing the underlying water quality issue. This is because a company could experience a greater number of cases of discoloured water but a smaller number of contacts, by virtue of contacting customers in advance more often than previously. According to the company in question, the research should therefore attempt to value the number of contacts directly, rather than attempt to map these contacts to the experience of the underlying water quality issues.

We do not agree with this proposal. Although companies may be able to reduce contacts without reducing the number of properties experiencing the issues, it is the water quality issues themselves that impact the customer primarily and not the experience of contacting the company. (This is a question that DWI may have a view on; however, it is customers themselves that should be the ultimate arbiter of what impacts upon them.)

When a customer is notified to expect an upcoming water quality issue, the impact of the issue is likely to be lessened. This is in the same way as we expect a notified supply interruption to have less of an impact than an unexpected interruption. However, not contacting the water company, having received a notification, does not imply a zero impact due to the water quality issue. Instead, this suggests that the PC measure itself has been defined more narrowly to focus solely on notified water quality issues.

In Appendix A, we set out an algebraic formulation of the mapping from experience of water quality issues to water quality contacts when the number of contacts depends on the number of notifications that are issued by the company, and where the impact on customers depends on whether or not they have been notified. This Appendix shows that the value to be assigned to the water quality contacts PC depends on the proportion of water quality issues that are notified rather than occurring at customers' properties unexpectedly. An estimate is therefore needed for this proportion in order to value the PC appropriately. Further details are given in the Appendix.

Application of the extended approach which considers the impact of notifying customers, will require additional input to set the following parameters:

- The likelihood that non-notified customers will contact the company about a water quality issue.
- The likelihood that notified customers will contact the company about a water quality issue.
- The proportion of water quality issues that are notified.

Sewer flooding

Expected PC definitions

Two sewer flooding measures have been proposed as common PCs. Definitions are subject to change, but could be as follows, based on PR19 definitions.

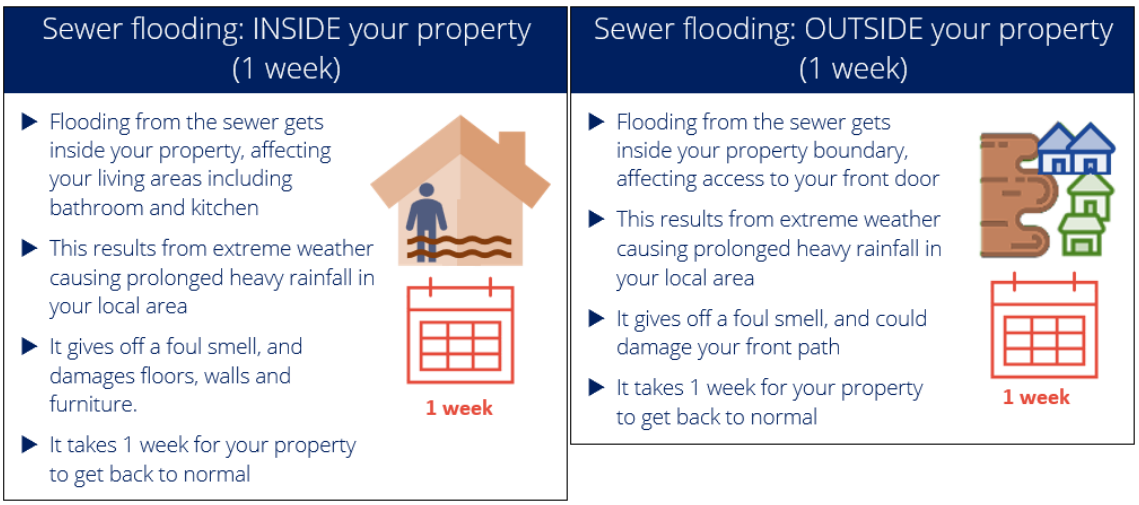
- Internal sewer flooding
  - The number of internal sewer flooding incidents normalised per 10,000 sewer connections including sewer flooding due to severe weather events.
- External sewer flooding
  - The number of properties affected by sewer flooding anywhere within the grounds of a building, unless it is already classed as an internal flooding incident. It does not include public roads and open spaces but does include sewer flooding due to severe weather events.

More detail around classification of sewer flooding incidents is given in Ofwat (2018).

Sewer flooding incidents can take a variety of forms and severities even within category. For example, interior flooding can range from a short-term minor toilet overflow to major damage being caused to property and health, and the customer needing to be re-housed for a long duration. Similarly, external flooding could affect only a small unused area of land within a property's boundary, or it could cover a large area and cause customers considerable harm. Accordingly, several companies have used 'Stage 2' stated preference research to derive impact weights for different types of flooding already, with the aim of providing a more nuanced valuation within their valuation frameworks.

For valuation, as with other measures, it is only necessary to include one type each of internal and external sewer flooding in the design. There may be benefits of including multiple types, but this brings an added complexity of needing to identify appropriate weights to place on the different types of incidents. As a minimum, we recommend that the survey design includes one type of internal and external sewer flooding only, that this is used for all companies, and that the type/severity is chosen to represent the median severities for the whole of England and Wales.

Figure 12: Sewer flooding issues definitions



Mapping
The only mapping needed in these cases would be to sum the number of incidents and, if scaled as in the PR19 common PC for internal sewer flooding, multiply by 10,000 and divide by the number of properties served

Environmental performance commitments

The second broad category of PCs to be considered contains those based on environmental impacts. This includes PCs supporting the following two outcomes:

- Sustainable use of water, including leakage, per-capita consumption (PCC) and business consumption.

- Environmental water quality, including pollution incidents, discharge compliance, storm overflows, Environmental Performance Assessment, bathing water quality and river water quality.

Although each outcome contains several individual PCs, they overlap with one another considerably and are, hence, appropriately treated jointly within groups with a view to optimally spanning the range of impacts therein whilst avoiding double counting.

Sustainable use of water

Performance commitment definitions

The PCs to be used to incentivise sustainable use of water are currently still under review (Ofwat, 2021). One option being considered is to include only a single PC based on distribution input; however, there could be three common PCs representing sustainable use of water, including:

- Leakage
  - The percentage reduction in the 3-year average leakage from the baseline period, where leakage is defined as the estimated loss of water from water company distribution assets (mains and service reservoirs) plus customer supply pipe leakage.
- Per capita consumption (PCC)
  - The percentage reduction in the 3-year average PCC from the baseline period, where PCC is defined as the sum of measured household consumption and unmeasured household consumption divided by the total household population.
- Business demand
  - No PC definition has yet been put forward, but we would expect a definition that parallels PCC, but is focussed on non-household customers rather than households.

Given that the PCs all measure water demand, it is necessary to consider how the demand for water impacts upon customers in order to define a customer focused measure. This has distinct differences depending on whether it is the short run, or the medium to long term, that is being considered.

In the short run, with fixed water available for use (WAFU), the level of demand, through any of the proposed PCs, determines the level of the supply-demand balance. This, in turn, determines the risk that demand restrictions will be required, including temporary use bans, non-essential use bans and emergency drought restrictions; as well as the risk of there being low flows in rivers.

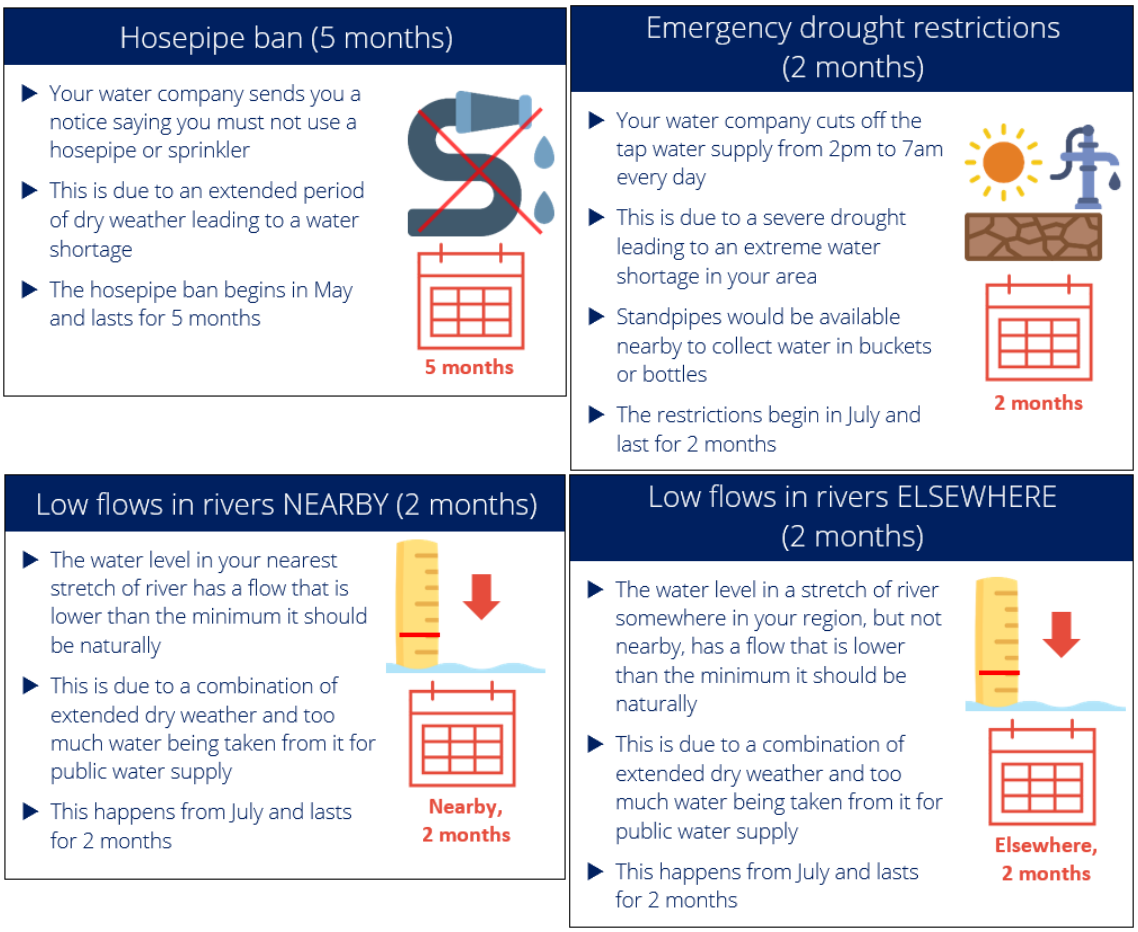
In the medium-to-long term, through the water resources management plan (WRMP) process, supply-side options could be developed or brought forward, which potentially mitigate against these impacts but at the cost of higher bills. For example, a company or

region) could potentially develop a new borehole, or reservoir or desalination plant, to compensate for lower levels of leakage or PCC reduction.

There are thus potentially different consequences associated with different demand levels depending on the time frame being considered. For the purposes of setting an ODI rate, it is the short-term impact that is appropriate for customer valuation. This is because ODIs are applied year-by-year, or on a 3-year average basis. Consequently, we propose that the survey should include customer measures based on drought restrictions and low flows.

Figure 13 sets out the proposed service issues to be used to capture the value per ML/Day of leakage and PCC. In line with other service issues, the definitions are precisely worded to avoid any ambiguity. Hence, they include the duration of the impact, its timing, its cause, and its location in the case of low flows.

Figure 13: Sustainable use of water issues definitions



Mapping

First, express Leakage, PCC, Business demand in ML/Day form. Then, determine the marginal impact on the expected number of days of each service issue per ML/Day of excess demand, i.e.:

$EDays_{TUB}$ = w_1ML/Day
$EDays_{NEUB}$ = w_2ML/Day
$EDays_{RotaCuts}$ = w_3ML/Day
$EDays_{LowFlow}$ = w_4ML/Day

Where:

$EDays_{TUB}$ is the number of days per year that a household customer is expected to be subjected to a temporary use ban

$EDays_{NEUB}$ is the number of days per year that a non-household customer is expected to be subjected to a non-essential use ban

$EDays_{RotaCuts}$ is the number of days per year that a customer is expected to be subjected to emergency drought restrictions

$EDays_{LowFlow}$ is the average number of days per year that customers can expect to experience excessively low flows in their nearest 1-mile stretch of river.

And:

$w_{i,}$ are the impacts on these quantities per ML/Day of excess demand.

NB: 'Expected days' is the probability weighted average of the number of days of restrictions, or low flow, that would occur given each possible weather/aridity scenario. Formally, this is given as $\int Days(a)f(a)da$, where *Days*(a) is the number of days of restrictions, or low flow, that would occur given aridity a, and $f(a)$ is the probability distribution of a.

Finally, the expected number of days of each type of service issue is to be valued by multiplying the expected number of days by the value per expected day, which is itself calculated as the value per avoided service issue (obtained from the survey) divided by the duration of the service issue as shown in the survey.

Under this construction, a 1-ML/Day increase in the leakage rate, for example, has a w_1 impact on the number of days per year that a household customer is expected to be subjected to a temporary use ban, a w_2 impact on the number of days per year that a non-household customer is expected to be subjected to a non-essential use ban, etc.

Application of the approach will require expert input to set the values of all of the w parameters linking 1ML/day variation around base distribution input to an expected number of days of each type of restrictions, and low flow. These are likely to vary by company or region, and this variation should ideally be taken into consideration.

A number of issues present themselves with respect to the proposed mapping:

1) The value per day of restrictions could be non-linear.

This means that choice of the duration levels to set for each of the service issues could impact estimates of the value per day of restrictions, and low flow. For the same reasons

as set out above in the case of other service issues, however, we would recommend that a median duration event should be used in each case.

2) It could be difficult to determine the impact of variation around base distribution input to the expected number of days of demand restrictions and low flows

In previous research for Thames Water we were able to determine this link for demand restrictions using data supplied by Thames Water and so do not expect that it should be prohibitively difficult. (See Metcalfe and Baker, 2011) We have not previously examined the impact on low flows, however, and so this aspect could potentially prove to be more difficult.

One option that could be considered to simplify the approach would be to focus on the risks of a standard restriction event, of fixed duration, rather than the expected number of days of restrictions. This approach would ignore the fact that more serious deficits lead to longer restrictions rather than simply a higher chance of restrictions of a fixed duration. The calculation may prove to be simpler, however, which could lead to this approach being preferable.

3) Leakage vs PCC incentives

The approach proposed results in equal incentives, per ML/Day, on leakage and PCC reduction. As suggested by Ofwat (2021c), however, it could be reasonable to expect that they may be different impacts on customer value due to the fact that reductions in consumption impact on customers, both positively through reductions in their bill, and negatively through the loss of the consumption itself. Reductions in leakage also potentially have external impacts, including increases in traffic when digging up roads to repair pipes, or through lower customer bills and/or customer disruption when repairing customer supply pipes.

In an ideal world, from an economic perspective, base service levels for PCC and leakage would be set optimally such that the marginal social cost of both, taking into account all externalities, was equal to the marginal value per ML/Day of water saved. In such a scenario, there should be no difference in valuations.

In practice, however, there are multiple distortions to the economic optimum in relation to PCC and leakage levels. With regard to PCC, the levels of consumption chosen by customers are potentially sub-optimal from a societal perspective due to the fact that prices are set to recover average costs rather than long-run marginal social costs, and because they are encouraged, on moral grounds, to choose consumption levels that may be sub-optimal to them, in terms of their own welfare, conditional on prices. With regard to leakage levels, companies are encouraged / required to undertake more leakage reduction than is implied by traditional cost-benefit calculations due, in part, to a belief that the costs and benefits estimated by companies do not adequately, or accurately, capture the true costs and benefits of leakage reduction, taking into account dynamic effects.

Overall, these factors support the idea that base levels of PCC and leakage reduction may, or may not be optimal, depending on how the distortions, and correctives, balance out against one another. If one assumes that base service levels are set optimally from an economic perspective, which must be the ultimate goal, then the assumption of equal

incentives, per ML/Day, between PCC and leakage reduction will still be appropriate. On this basis, we recommend that the approach proposed here is used, which involves equating the incentives per ML/Day between PCC and leakage reduction.

4) Use of 'Nearest stretch of river'

Whilst all issues thus far have focused on service issues at customers' properties, the proposal here now includes an issue that affects the environment: a period of low flows at the customer's nearest 1-mile stretch of river. This introduces a new set of issues due to the fact that environmental impacts affect multiple customers, and in a somewhat different manner in comparison to private service impacts.

As is discussed extensively in the literature on environmental valuation (see, for example, Freeman, 2003, and Turner, 1999), values for environmental protection or enhancement derive from a variety of motivations. For example, when I say I am willing to pay for the prevention of harmful impacts to a river system, this could reflect a variety of motivational beliefs:

1) It may affect my enjoyment of future recreational visits to the river in question.
2) The environment should be looked after for its own sake and/or for the sake of wildlife (and plants) that depend on it.
3) It may affect others' enjoyment of future recreational visits to the river in question, which I care about because:
    a) I care directly about the environmental quality experienced by these other people during these visits; or
    b) I care about these other people's wellbeing including, but not limited to, their enjoyment of the environment.

Beliefs (1), (2) and (3a) reflect valid values that should be counted, while Belief (3b) should not be included in the sum of values as this would involve double-counting due to the fact that each person's own value is already counted.

Belief 3), often termed 'existence value', represents values that are not based on anyone's use of the river. The question of whether such values should be incorporated within societal valuations has been debated in the literature (see Rosenthal and Nelson, 1992, and Kopp, 1992). The general consensus is that they ought to be included as they are part of Total Economic Value (Bateman et al 2002; Pearce et al. 2006). However, there remain debates around the ability of stated preference methods to measure such values reliably. (See Hausman, 2012, and Carson, 2012 for example.)

The key difficulty in obtaining valid values that incorporate existence value motivation is that such values tend to be inadequately sensitive to scope. Again, there is an extensive academic literature on the sensitivity to scope of environmental valuations, with mixed results. (See Lopes and Kipperberg, 2020, and Burrows et al., 2017, for recent papers on the topic.)

When specifying the environmental impacts to be included in the survey for the Collaborative ODI research, we have been mindful of the potential for scope insensitivity to confound the estimation of reliable valuations. We considered two options for the specification of impacts on rivers:

- Option 1) Focus on the nearest stretch of river only.
- Option 2) Include low flows in a stretch of river far away from the customer's property as an additional service issue

The advantages of Option 1) are:

- it is easier to imagine your nearest stretch of river than it is to imagine a stretch of river far away
- it presents an issue that is as comparable as possible to issues at the customer's property, which thus facilitates comparison of impacts
- it results in a conservative valuation as it would implicitly assume that values for protecting non-nearest rivers are equal to zero. Whilst this assumption could be challenged, it provides an offset against the counteracting tendency for participants to overstate the impact on them from harm to rivers due to them caring about the welfare of other people. As previously discussed, this value should not be counted because it is already counted elsewhere via the method of summing each individual customer's estimated value.

The key advantage of Option 2) is, by contrast, that it avoids implicitly assuming that values for protecting non-nearest rivers are equal to zero.

On balance, in the interests of ensuring flexibility, we recommend focussing on Option 2), and have specified the proposed service issues accordingly. This option does not preclude assigning the river that is far away a weight of zero as a conservative approach, whilst it leaves open the option of assigning it a positive weight.

Environmental water quality

Performance commitment definitions

The PCs to be used to incentivise environmental water quality protection and enhancement are currently also under review (Ofwat, 2021c). At PR19, common PCs in this area included only pollution incidents and discharge compliance. For PR24, Ofwat is considering adding bathing water quality, river water quality and storm overflows. Ofwat is also considering whether and how to use the Environment Agency / Natural Resources Wales Environmental Performance Assessment (EPA) rating as a PC.

For 2021, the EPA is an overall assessment of 6 metrics that set out how the 11 water and wastewater companies comply with specific obligations that environmental regulators enforce. These metrics are:

- Discharge compliance
- Total pollution incidents
- Serious pollution incidents
- Proportion of pollution incidents that are self-reported
- WINEP scheme delivery
- Supply Demand Balance Index

For the 2026-2030 period, the Environment Agency is considering adding a metric for storm overflows to this list.

Clearly, there is an overlap between EPA and the individual metrics proposed for inclusion. Ofwat is considering whether the EPA is needed in addition to metrics for discharge compliance, pollution incidents and storm overflows, or whether it might replace these metrics. In either case, bathing water quality and river water quality are proposed to be included within the suite of PCs for PR24.

Given the uncertainty over which PCs will be included, we discuss mapping at a higher level in the following than previous parts of this section.

Figure 14 sets out the proposed service issues to be used to capture the value of impacts on environmental water quality. The definitions again include the duration of the impact, its timing, its cause, and its location.

For the same reasons as discussed above in the context of sustainable use of water service impacts, the river impacts include both nearest river and far away river. As a conservative approach, the value associated with the far away river could be discounted; however, including these within the set of service issues explored allows for greater flexibility.

Figure 14: Environmental water quality issues definitions

Pollution incident NEARBY (2 days)

- Untreated sewage spills into your **nearest** stretch of river
- This is due to heavy rainfall in your local area
- There would be minor visible pollution and damage to the river, with 10 fish dying
- The spill begins on a Wednesday and lasts for 2 days. The river is then back to normal after 1 week

Nearby, 2 days

Pollution incident ELSEWHERE (2 days)

- Untreated sewage spills into a stretch of river somewhere in your region, but not nearby
- This is due to heavy rainfall in that area
- There would be minor visible pollution and damage to the river, with 10 fish dying
- The spill begins on a Wednesday and lasts for 2 days. The river is then back to normal after 1 week

Elsewhere, 2 days

River water quality NEARBY less than 'Good Ecological Status'

- Your nearest stretch of river is consistently at less than Good Ecological Status, as defined by the Environment Agency, although it does achieve Moderate Ecological Status ⓘ
- This is due to a variety of factors, including the quality of treated wastewater, the river flow level, and the run-off from the surrounding area ⓘ
- There are fewer and smaller fish than there would have been, and occasional algal blooms ⓘ

Local

River water quality ELSEWHERE less than 'Good Ecological Status'

- A stretch of river **in your region, but** not nearby, is consistently at less than Good Ecological Status, as defined by the Environment Agency, although it does achieve Moderate Ecological Status ⓘ
- This is due to a variety of factors, including the quality of treated wastewater, the river flow level, and the run-off from the surrounding area ⓘ
- There are fewer and smaller fish than there would have been, and occasional algal blooms ⓘ

Elsewhere

Coastal bathing water quality nearby less than 'Excellent Status'

- The sea water at nearest beach is consistently at less than Excellent Status, as defined by the Environment Agency, although it does achieve Good Status ⓘ
- This is due to the quality of treated wastewater entering the water nearby
- You could still swim in the sea, but there would be a small increase in the chance that you might get ill if you swallowed some water ⓘ

Water less than excellent

Mapping

Pollution incidents and storm overflows

1 pollution incident/storm overflow is valued by multiplying the average customer value per incident by the number of customers whose nearest stretch of river is affected. The average customer value per incident is derived from the survey responses; the number of customers whose nearest stretch of river is affected is calculated by multiplying the total number of customers supplied by the company by the ratio of the number of miles affected by the incident to the total number of river miles in the company area.

For example, if there are 1 million customers and 500 miles of river in the company area, and a pollution incident affects 1 mile, then the value per pollution incident would be calculated as 2,000V, where V is the value per customer affected and 2,000=1 million/500.

Discharge compliance and river water quality

These PCs appear to overlap and, as such, the total value for the two PCs needs to avoid double counting. The approach set out in Appendix A should be used to determine the balance between the two PCs. It shows that in order to value overlapping PC measures, it is necessary and sufficient to agree the impacts of a 1-unit change in each PC, on the likelihoods of each underlying customer-facing issue. In the present case, river water quality is one of the PC measures as well as the sole relevant customer facing issue. In order to justify the inclusion of discharge compliance as a PC in addition to river water quality, there must be some informational distortion in the river water quality PC measure that can be alleviated by discharge compliance. Otherwise, there would be no additional customer value from improvements in discharge compliance over and above their impact on river water quality. If, however, the informational distortion implies that there is an additional customer value from measured improvements in discharge compliance over and above measured changes in river water quality, then the approach in Appendix A can be used with two PC measures and one customer issue, treating the customer issue (river water quality) as distinct from the river water quality PC measure for the purposes of valuation.

Bathing water quality

1 bathing water improvement to Excellent from Good is valued by multiplying the average customer value per bathing water falling short of Excellent by the number of customers whose nearest bathing water is affected. The number of customers whose nearest bathing water is affected is calculated by dividing the total number of customers supplied by the company by the total number of bathing waters in the company area.

Recommended service issues for compensation exercise

For the compensation exercise, two service issues are recommended to be included. Only one is needed to obtain monetary values for all of the service issues included in the

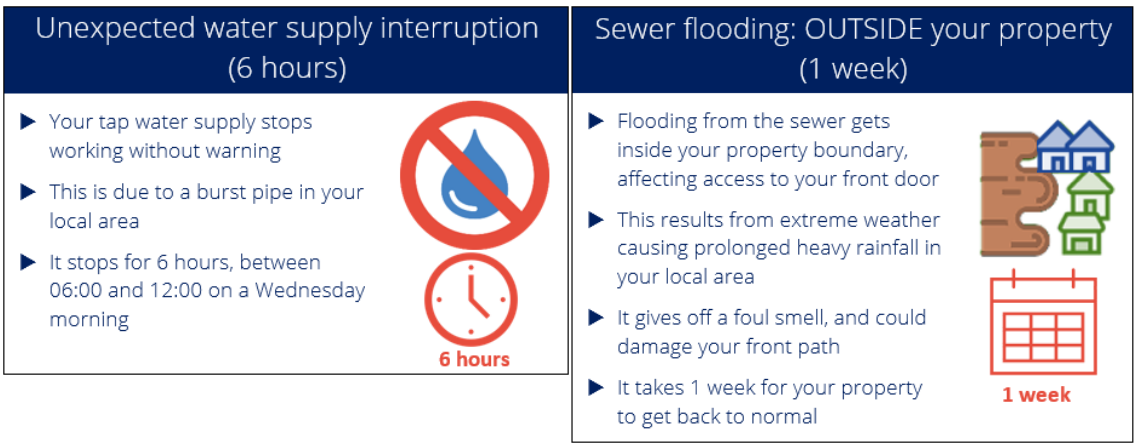
impact-based choice exercise; however, two are recommended in order to mitigate against any over-reliance on a single estimate.

In selecting which service issues to include, we have been guided by the following principles:

- Ideally, one water service issue and one wastewater service issue should be included. Whilst not required by the methodology, this would allow valuations for each service type to be pivoted off a valuation for a service issue within the same area.
- The service issues should be ones that impact customers directly, rather than impacting their local environment. This is for two reasons:
  - Compensation is likely to be perceived as significantly more credible in the case of service issues affecting customers' properties directly than for impacts affecting their environment.
  - The impacts themselves are less abstract, and hence likely to be more reliably measured for service issues directly affecting customers' properties.
- Finally, service issues with a lower required compensation are likely to work better. This is because the compensation levels required for people to be indifferent, for example, between experiencing internal sewer flooding and not experiencing it are likely to be so large as to appear incredible to participants. This could lead to them not accepting the scenarios at face value, and hence not giving meaningful responses.

In line with these principles, the two recommended service issues are a short, unexpected supply interruption, and an external sewer flooding incident. These are shown in Figure 15 below.

Figure 15: Service issues to be included in compensation-based exercise



3.5 Choice formats

Impact-based exercise

Figure 16 presents the proposed introduction to the impact-based exercise, for households. It includes a note explaining how 'impact' should be interpreted in a form that is consistent with economic valuation. This is particularly significant in relation to the environmental impacts as the valuations here are intended to reflect Total Economic Value, which includes non-use value, but should not include general altruism towards other people as this would lead to double counting.

Figure 16: Introduction to impact-based exercise (household version)

You are now going to be shown a series of questions each giving two different scenarios related to your water or wastewater service. In each case, please consider both scenarios carefully and choose which of them would have the most impact on your household.

Some of the service failures shown would affect your own property whereas others would affect your local area. When comparing the impact that each would have, please:

- do consider any concerns you may have for the local area or environment,
- don't consider any impacts on other people outside your household - other people will answer for themselves!

Figure 17 and Figure 18 show examples of the kind of questions envisaged for this exercise for landscape screens, e.g., desktop computers or laptops, and portrait screens, e.g., smartphones, respectively. In contrast to the MaxDiff version of the exercise reviewed in Section 3.2, the choices are presented in pairs rather than groups of four. This approach was taken in order to be able to display all the required information on the screen at the same time without requiring participants to click on buttons for further information, which can be awkward, and hence not often done, when using a smartphone to complete the survey.

Figure 17: Example impact-based exercise question (household version, landscape format)

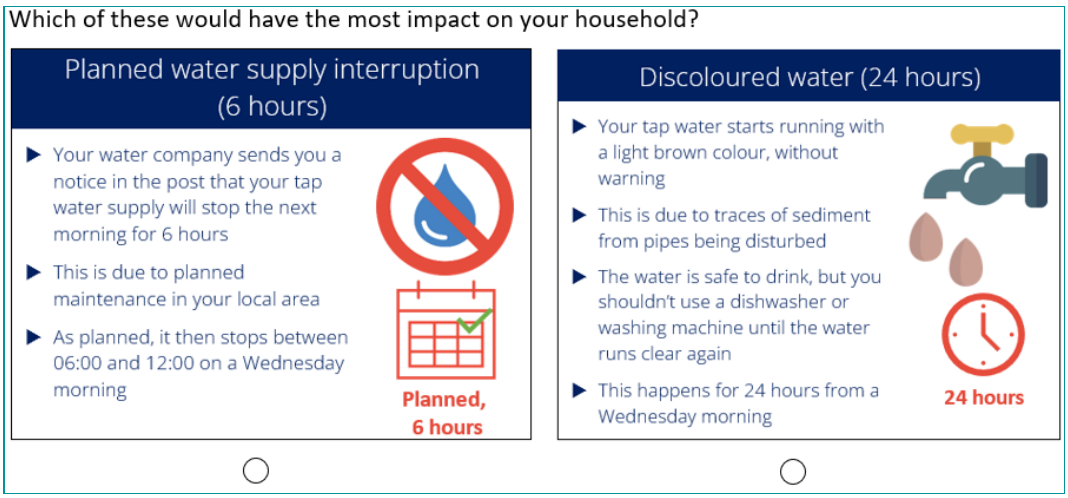
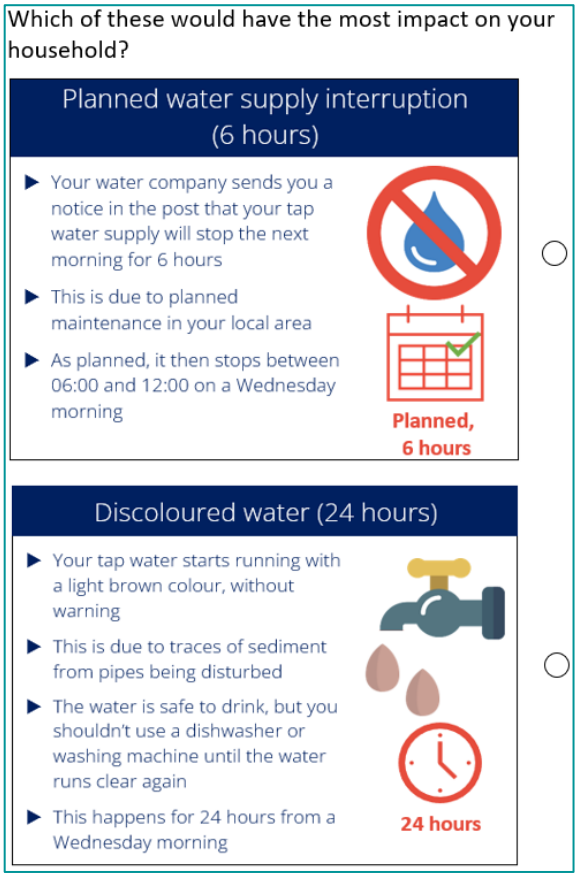


Figure 18: Example impact-based exercise question (household version, portrait format)



The above examples are tailored to households. However, the non-household versions would be almost identical. The only difference envisaged is that the question would instead ask: 'Which of these would have the most impact on your organisation?', rather than 'on your household'.

Compensation-based exercise

Figure 19 presents the proposed introduction to the compensation-based exercise, for both households and non-households. The key purpose of this introduction is to provide a credible scenario for the choice being asked of participants, which should then encourage them to answer carefully.

Figure 19: Introduction to compensation-based exercise

Water and wastewater companies currently pay their customers compensation in some cases when there are problems with their service. They also invest money to reduce the number of problems that happen in the first place.

There is now a choice between:

- spending more to reduce the number of service problems
- paying out more compensation to customers that experience the problem.

To help decide this, the next few questions will each present you with a choice between experiencing a service issue and being compensated, or not experiencing the issue and not receiving any compensation.

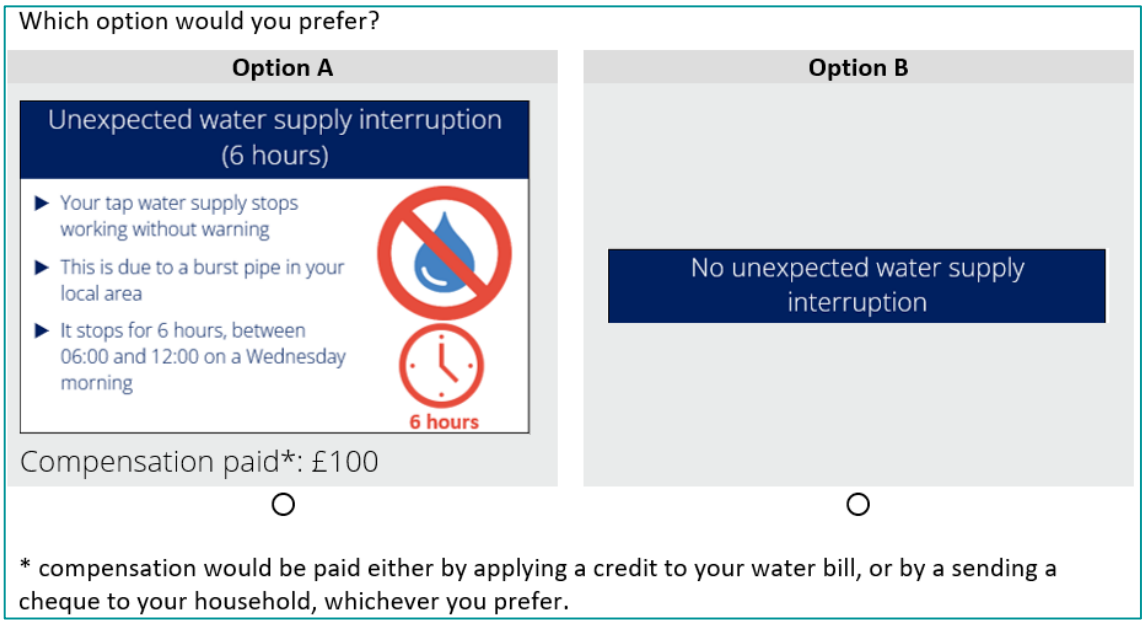
In each question, the type of service problem and the compensation amount will vary.

Your answers to these questions will help decide where to strike the balance between the levels of compensation paid and the number of service problems experienced by customers.

Figure 20 shows an example of the kind of question that would be asked in this exercise. A landscape format version is shown in this figure. However, a portrait version will be shown to participants answering on portrait-oriented screen (e.g. smart phone), as for the impact-based choice questions.

In this exercise, the compensation paid would vary across participants using amounts selected to maximise the statistical efficiency of the valuations ultimately obtained. A note is included to say how the compensation would be paid. It is important that participants view the payment mechanism as credible, and so this will be tested in Stage 2 of the study.

Figure 20: Example compensation-based exercise question (household version, landscape format)



The above example is tailored to households. The non-household version would be similar. However, we recommend that compensation levels are specified as percentages/multiples of the annual bill in this case due to the fact that bills vary enormously across non-households. This recommendation is consistent with approaches typically taken in water and wastewater service valuation surveys, and in line with the UKWIR 2011 guidelines (NERA-Accent, 2011). Additionally, the footnote would substitute 'organisation' for 'household' when explaining how compensation would be paid.

3.6 Experimental design

Once the service issues have been agreed, an experimental design is needed for both exercises in order to create the choice situations that participants will see. This requires first selecting which service issues are included in surveys, i.e. whether water and wastewater service issues should be grouped or kept separate, and then creating the design that combines them into choice situations.

Grouping of water and wastewater service issues

An important question to be addressed before creating these choice situations, which affects survey administration and sampling as well as design, is whether water and wastewater service issues should be grouped together or separated. The advantage of grouping them together is that the sample size would be effectively doubled for the same number of participants, in comparison with keeping them separate. On the other hand, two water only companies raised the issue within the industry consultation about whether the joint inclusion of wastewater service areas would affect the water service valuations relevant to their company.

Our recommendation, in response to this question, is that the survey should cover both water and wastewater service issues for every participant, regardless of whether they are supplied by the same company for both water and wastewater services, or whether they have separate suppliers.

The key advantage of grouping them together is that the sample size would be effectively doubled for the same number of participants. An additional advantage is that it leads to greater comparability across the industry.

These advantages are gained with no loss to water only companies in terms of their valuations being affected by the inclusion of wastewater services. This is because valuations are 'pivoted' off the value for one service issue derived from the compensation exercise rather than being dependent on a package valuation.

Impact-based exercise design

For the impact-based exercise our recommended methodology for creating the designs involves applying the 'D-efficiency' design approach (Rose and Bliemer, 2009). This requires the specification of prior values for the impact weights ultimately to be derived. The approach then uses these to calibrate the selection of service issues so as to maximise the statistical precision of the estimates ultimately obtained.

We propose to use the PR19 WTP values contained in Accent-PJM Economics (2018b) to set priors initially. For the main stage, estimates from the pilot survey analysis should be used to re-calibrate the experimental design to improve its statistical efficiency.

An alternative approach often applied, but which we do not recommend, would be to create a balanced incomplete block design where each service issue appears an equal number of times, and with each other service issue an equal number of times. This will,

in general, result in weaker statistical estimates than the recommended approach as it is not optimised for statistical efficiency. The recommended approach achieves a greater level of statistical efficiency, via calibration, by tending to place service issues alongside one another that are close together in terms of predicted impact. This makes for more meaningful choices as well as providing for more precise estimates.

In creating the experimental design, the approach allows any number of choice questions to be asked of each participant, and any number of distinct sequences of choice questions to be allocated across the sample as a whole. With regard to the number of choices, more means better in terms of statistical precision. However, there is a limit to the number that each person can reasonably be expected to answer before showing signs of fatigue. There are no hard and fast rules for what number should be chosen as it depends on the complexity of the choices, and of the survey as a whole. For the initial phase of testing, we recommend including 10 questions per interview. This may be reduced to 8 if there are indications in the cognitive testing phase that 10 is too many.

With regard to the number of distinct sequences of choices, or blocks, there are typically gains to be made to statistical efficiency from having multiple blocks rather than just one, and also advantages in terms of mitigating any ordering effects due to the exact sequence of choices. These gains diminish to zero as the number of blocks increases. When creating the experimental design, we would therefore compare different-sized designs and select one with an appropriate number of distinct blocks.

Compensation-based exercise design

For the compensation-based exercise, there are two service issues included, and these are to be the same for each survey participant. The only variables are, therefore, the order in which the service issues are asked; the number of different compensation amounts asked about for each service issue; and the amounts of compensation shown on each occasion.

To mitigate against the impact of order effects, we recommend randomising the order of the service issues that appear in the compensation-based choice exercise, with half the sample seeing the supply interruption question first and the other half seeing the external sewer flooding question first. We also recommend randomising the range of compensation amounts that are shown. Initially, this will cover a broad range, based on findings from Accent-PJM Economics (2018a), but will be refined following the pilot to optimise the statistical precision of the design.

We propose that two questions are included for each service issue, thus meaning there would be four questions in total.

The amount of compensation shown in the second question for each service issue would depend on the answer to the first question, in line with the double-bounded contingent valuation design methodology. This means that if the participant chooses 'no interruption', for example, at the first question, the compensation amount offered would be doubled in the second question. Alternatively, if they choose 'interruption plus compensation' at the first question, the compensation amount offered would be halved in the second question.

As has been established in many studies within the stated preference literature (e.g. Boyle et al., 1985; Herriges and Shogren, 1996; Whitehead, 2002; Day et al., 2012), valuation tasks can be subject to ordering effects. Different explanations appear for these effects in the literature, which range from them being the outcome of strategic choice based on the perceived incentive properties of the choice questions (Carson and Groves, 2007) to them being the outcome of on-the-fly preference construction (e.g. Green et al., 1998; Ariely et al. 2003).

Accordingly, we anticipate that the compensation amount shown in the first question could have an anchoring effect on subsequent responses and, for this reason, the first question shown takes a particular prominence. The analysis of the data arising from this choice exercise will therefore need to be sensitive to the possibility of these effects and use a modelling methodology capable of handling them appropriately.

3.7 Issues concerning the value measure obtained

As part of the industry consultation undertaken for the study, a number of questions were raised regarding the nature of the valuations that ought to be obtained, and that will be obtained under the recommended approach. Most of these questions have been addressed elsewhere in the document, but two remaining questions are collected and addressed in the following text.

Is there a risk of over-valuation due to the use of a WTA measure of value in the compensation-based approach, rather than a WTP measure?

In most previous valuation research in the water sector, evidence on customer values of marginal changes in service levels has been derived by examining customers' willingness to pay for improvements in service levels. Where WTA values and WTP values have both been estimated, (e.g. Lanz et al. 2010), it has been found that WTA values exceed WTP values for the same service increment. This finding is consistent with the broader literature in which it has been very commonly observed that WTA exceeds WTP.

The approach recommended here derives values based on customers' willingness to accept compensation for service failures that impact on them. It is therefore a natural question in this context as to whether this is likely to lead to an over-valuation in comparison to previous approaches.

Although a WTA approach is being used in the present study, the bigger difference, in terms of anticipated impact on the size of the valuation results, is that valuations are now being derived via consideration of the utility difference between experiencing and not experiencing a service issue, rather than via the utility difference between a small risk of experiencing a service issue and a slightly smaller risk of experiencing the same service issue.

It would be possible, in principle, to apply the same broad approach in a WTP context by asking customers if they would be willing to pay to avoid an otherwise-certain service issue impacting upon them. This would almost certainly return lower valuations than those likely to result from the proposed WTA approach. However, this approach would

be unnatural, and liable to be perceived as incredible. Moreover, the resulting valuation, whilst certainly more conservative, would not be the correct valuation to use due to the fact that customers can expect, on a day-to-day basis, that the default position is one where nothing goes wrong, rather than one where they experience failure after failure. It is hence correct to use a WTA measure in this context, as well as being a more practical approach that is more likely to work well with customers in a survey context.

Are there differences between customer preferences as individuals and as citizens, and how are these taken into account?

One company raised the question as to whether it should be customer preferences as citizens that should be measured rather than customer preferences as individuals, which it is supposed are the focus under the proposed approach. This is a deep question, which cannot be fully addressed here.

Cost-benefit analysis as a method for public decision making has been criticised for its reliance on the aggregation of individual's preferences, as opposed to relying on deliberative public reasoning. (See Orr, 2007, for a discussion).

However, within the framework of CBA, the distinction between individual preferences and citizen preferences can be understood to relate to the question of whether altruistic values ought to be included, which is a question that has been addressed exhaustively in the CBA literature. Bergstrom (1982) and McConnell (1997), for instance, demonstrated that including values based on general concern for others' welfare leads to double counting and should be avoided in a social CBA. This view is generally accepted among practitioners; see for example Bateman et al (2002).

It is therefore appropriate for values that are to be used in cost-benefit appraisals, and by implication ODI rates, to be based on individuals' own preferences, and that these should be summed over the full population. This does not rule out concerns for the environment, and the wildlife that depend on it, being a factor in individuals' preferences, but the survey should encourage participants not to consider the impacts of service issues on other people when considering how much impact each type of service issue will have on them, as this would result in double counting.

3.8 Questionnaire structure

In line with commonly applied standards for stated preference research, (e.g. Bateman et al., 2002, p.148-151), the recommended questionnaire structure for the Collaborative ODI research includes the following components, set out in Table 7.

Table 7: Questionnaire structure

Questionnaire component	Purpose
Screening	Ensure the sample conforms to design using quotas
Introduction	Introduce the survey
Experiences and uses	Provide data expected to correlate with stated preference responses as a check on their validity
Attitudes	
Impact-based exercise	Core stated preference data
Follow-ups	Check validity of responses
Compensation-based exercise	Core stated preference data
Follow-ups	Check validity of responses
Socio-demographics	Provide data expected to correlate with stated preference responses as a check on their validity, and provide any further data for segmentation, beyond what is included for screening.

This structure is focused on the core stated preference exercises only, with additional material included to: ensure the sample conforms to design; introduce the survey and its components; check and verify the validity of the stated preference responses; and provide data for segmentation.

4 Survey administration and sample design

4.1 Introduction

This section presents options for various aspects of survey administration and sample design, the advantages and disadvantages of each option, and our recommendations of how this is progressed to obtain an optimal main survey approach.

It also presents the results of an online workshop with stakeholders on 13 December 2021. In this workshop, options for each aspect of survey administration and sample design were first presented in a plenary session, and then discussed among participants in breakout groups. Participants also chose their preferred option in two surveys, before and after the breakout group discussions.

There has been a substantive amount of subsequent discussion amongst the delivery team to explore the detailed nuances of all approaches and this is also reflected in this section.

The following options are covered in this section:

- Household survey method: A choice between commercial online panels potentially supplemented with face-to-face recruitment/surveying of hard-to-reach customers and drawing a random sample of customer postal/email addresses from company billing files or from the Postal Address File
- Non-household survey method: a choice between using commercial online panels with screening questions to filter in those that have responsibility for paying utility bills for their company, a Postcard drop to a random sample of postcodes, email and/or phoning customers via lists obtained from retailers and phoning businesses at random (from a sampling frame such as Dun & Bradstreet) and with both phone approaches utilising a phone-email/post-phone method
- Sample sizes and segmentations: a choice between different sample size options with larger companies having larger sample sizes compared with an approach based on integrating wastewater and water issues together for every customer, including those where services are supplied by separate companies, with proportional sample sizes used as the basis for matching the customer base.

4.2 Household survey method

Below we set out the two options which were presented at the December workshop and set out their strengths and weaknesses, stakeholders views and our subsequent recommendations based on further detailed discussion.

Options

Option 1: Using commercial online panels (90%), supplemented with face-to-face recruitment and surveying of hard-to-reach customers (10%)

The potential advantages of this option are:

- Online panels are cost effective, quick and return samples that are potentially representative on demographics through the use of quotas (age, gender, Socio-economic group, region, urban/rural).
- It is a widely used and well understood methodology in both the water sector and more broadly.
- Panel participants are used to taking surveys and are consequently potentially more adept at understanding survey instructions and working through a series of questions.
- The face-to-face component is typically designed to mitigate against the potential omission of digitally excluded households.

The potential disadvantages are:

- The panels are typically felt to need to be supplemented by face-to-face interviews to ensure good coverage as described above, which potentially conflates mode and demographic effects in the sample

  The panel participants are not necessarily representative on other non-observed measures:

  - They may be more cost sensitive because they regularly give up lots of time to complete surveys for relatively small financial reward.
  - They may be more computer savvy than average
  - They may have been recruited based on particular unknown characteristics (e.g., Nectar card users), and this can be seen to be something of 'a black box'
  - Any inherent such issues/potential biases are not necessarily consistent between water company areas or between panels
- The quality of response may be affected by survey fatigue as participants undertake a number of surveys on a regular basis
- The numbers of people on such commercial panels are limited in the smallest company areas – hence feasible sample sizes may be smaller than ideal, or require a different mix of research methods to bolster the shortfall which could lead to concerns over inter-company comparability.

Option 2: Draw random sample of customer postal/email addresses from billing (or alternative sampling) files

In this approach customers are typically contacted by their preferred method of contact (post/email/phone) and invited to complete the survey. The survey itself can be completed online or by post to provide opportunity for all.

The invitation would be sent to the occupier of the household and allow anyone to respond that self-certified as being able to answer on behalf of their household. This approach was considered preferable to adopting a randomised approach to sampling within household because the key success factor is that the recruited sample member is able to answer on behalf of their household.

The potential advantages of this option if company customer lists are used as the sampling basis are:

- The sampling frame can have near complete population coverage, i.e., there is for some companies a good match between sampling frame and population of interest
- The starting point is a potential random probability sample
- A full sample can be available for small companies
- Stratification is possible (e.g., according to area, social tariff, use of meter)
- Good contact details are typically held for those on social tariffs (e.g., telephone numbers and/or email addresses)
- Multi-mode options (e.g., phone/email/telephone) are possible though the main options being considered are postal and online
- It is possible to link to other customer data (e.g., bill amounts), either for use in the survey itself or for subsequent analytical purposes
- There is potentially more consistency across companies.

The potential disadvantages are:

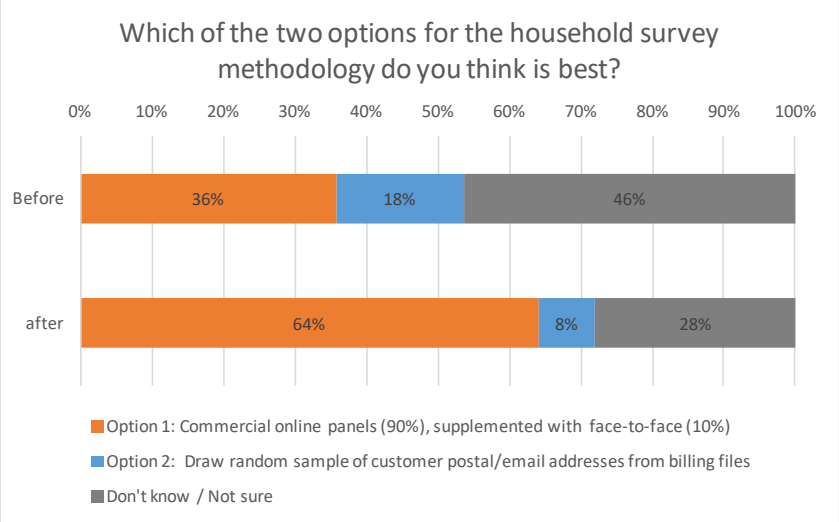
- It can be more expensive (incentives, mailouts, design flexibility to enable the use of different modes of survey delivery, liaison with companies)
- It is potentially more burdensome on companies (drawing sample, handling GDPR issues)
- It does not include non-bill payers
- Bill payers may be landlords covering many properties rather than households
- Certain segments, e.g., younger people, may be less likely to respond. Hence, a larger sample size may be needed to achieve the same levels of precision in the results

- Response rates could be low unless relatively large incentives are used and reminders sent to non-responders, which would increase the time and cost of the survey.

Stakeholder views

The figure below shows the views of workshop participants about the household survey methodologies. Option 1 (commercial online panels) was the preferred option, chosen by 36% of participants before the discussions and by 64% after the discussions.

Figure 21: Stakeholder views about household survey methodologies



The views set out below reflect the balance of opinion at that time on these options.

During the breakout workshops, commercial online panels (Option 1) were seen as a pragmatic approach, with several potential advantages cited:

- Good response rates
- People in the panels get used to answering surveys, so "they know what they are doing" though this could also be presented as a concern that participants come with too much knowledge
- Online panels have been used by water companies for some time, so they bring known risks
- The 10% face-to-face boost is also a practice some water companies are also used to including - it is commonly used to reach hard to reach and digitally excluded populations.

By comparison, Option 2 (using customer lists) was seen to have several disadvantages by the workshop participants to sit alongside the potential benefits of better sample representation:

- The risks of this option are less well known by water companies, compared with using online panels.

- The response rate may be low leading to potential non response biases.
- There are potentially data privacy issues and regulations on data sharing – there will be a potential need to reassure customers about where their contact data came from.
- Extra costs, environmental impacts and effort of mass mail outs.
- It introduces another stage in the survey, leading to some customers giving up answering after they receive the invitation and potentially affecting hit rates.
- The postal sampling approach may potentially skew the sample with, for example, a higher proportion of older customers answering than younger customers.
- It is not clear how to reach people who are not the bill payers with this approach.
- Databases of email addresses and phone numbers are not always complete. Customer data varies from company to company and customer to customer which could lead to issues of comparability.
- The customers' preferred means of contact may not be up-to-date.
- Comparability issue of surveys undertaken with different recruitment and survey modes.

Subsequent development discussions and recommendations

Subsequent discussion between delivery team members led to the following recommendations:

- That both approaches should be developed and piloted to explore issues of practicality, efficacy and cost and reflecting the strengths of opinion amongst supporters of both approaches.
- That company customer contact databases be replaced with the use of the Postal Address File to broaden coverage to include non bill payers; the PAF approach was also felt to address the potential comparability issues associated with varying company database qualities as well as mitigating some of the delivery risks.
- That the questionnaire would not be sent to potential participants in the initial PAF approach mail out to minimise any environmental issues associated with mass mailing.
- That the commercial panel approach would not also include a face to face element at this stage but that further consideration would be given to the digitally excluded if the commercial panel approach was the ultimate recommendation.
- That landlords would be excluded through screening questions.
- That differing incentive regimes would be tested to look at the impact on hit rates for the PAF approach.

4.3 Non-household survey

Below we set out the options that have been under scrutiny and their strengths and weaknesses, stakeholders views and then our recommendations.

Options

Option 1: Commercial online panels – screening questions used to filter in those that have responsibility for paying utility bills for their company

The advantages of this option are:

- Online panels are relatively cheap, quick and have tended to return samples that are fairly representative by size band
- It is a widely used methodology in water and in other sectors.

The disadvantages are:

- There is the risk that many participants do not actually make decisions about their company's utility bills
- Panel participants are not necessarily representative on other non-observed measures.
- Numbers of people on panels are limited in the smallest company areas and hence feasible comparable sample sizes may be smaller than ideal.

Option 2: Postcard drop to random sample of postcodes

The advantage of this option is that it is possible to move towards each address having a reasonably equal probability of being contacted.

The disadvantages are:

- Certain segments may be less likely to respond to a postcard drop leading to potential biases
- Response rates could be low unless incentives are used and follow-up contacts are made, which would increase the time and cost of the survey
- It is not a tried and tested methodology – its usage has been limited
- When sampling it would be normal for cost reasons to have this clustered which would undermine attempts at a random sampling approach.

Option 3: Email and/or phone customers via lists obtained from retailers

The advantage of this option is that it should theoretically be possible to ensure that each customer has an equal probability of being contacted.

The main disadvantage is, on the other hand, that the option may not be possible for all retailers which would create clusters of database availability by company given the way that the market is structured.

The potential use of email contact details for some participants would provide potential cost savings though it does mean that two modes of approach – email and phone – are being used which could have comparability concerns.

With this option, there is greater confidence that the person completing the survey is the correct decision maker.

Option 4: Phone businesses at random (from, for example, a Dun & Bradstreet sampling frame), then post/email materials, and either carry on with the interview or book an appointment to call back to complete the survey (the phone-email/post-phone approach)

With this option, there is also greater confidence that the person completing the survey is the correct decision maker.

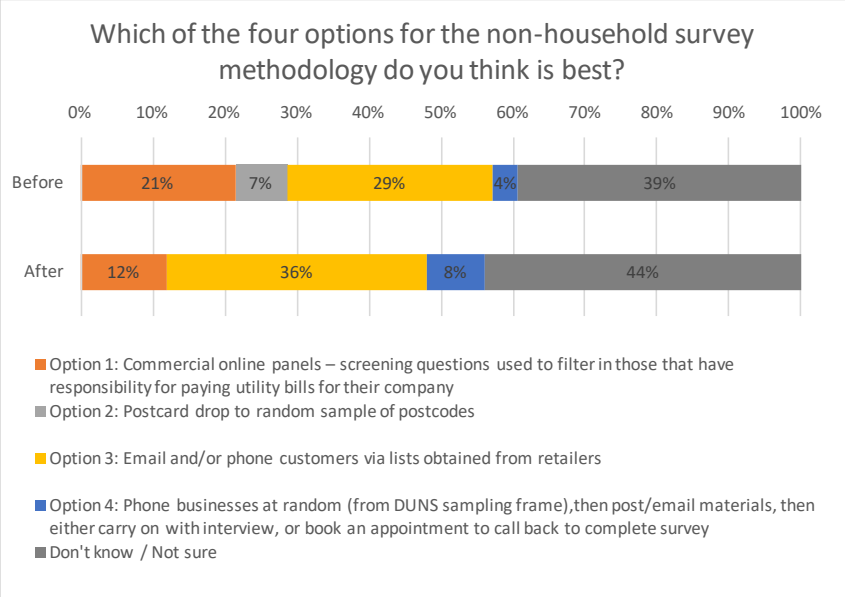
But it does come at a cost as each landed interview requires a number of unsuccessful recruitment attempts which we would expect to be higher than when using the retailer sample.

It is a tried and tested methodology in the water and other sectors.

Stakeholder views

Figure 22 shows the views of workshop participants about non-household survey methodologies. Option 3 (email and/or phone customers via lists obtained from retailers) was the preferred option, chosen by 29% of participants before the discussions and by 36% after the discussions. The second most preferred option was Option 1 (commercial online panels), chosen by 21% of participants before the discussions and by 12% after the discussions.

Figure 22: Stakeholder views about non-household survey methodologies



In the workshop discussions, participants emphasised the importance of reaching the relevant person in the business, i.e., the one who makes the decision about utilities.

This does count against the use of commercial panels to a degree though it is something to be wary of with all NHH approaches.

In general, it is difficult to get people interested in filling lengthy surveys, even with incentives.

Reaching businesses by phone is preferred to a postcard drop based approach.

Subsequent development discussions and recommendations

The preferred approach centres on Options 3 and 4.

The core approach we recommend being piloted is a phone-email/post-phone one and the pilot testing would be to see whether the seemingly simpler questionnaire leads to an improvement in hit rates when compared with the PR19 work.

Discussions also continue with the retailers to see how many would be willing and able to provide assistance. The potential gain from this would be the potential availability of email addresses and named individuals for some which would have a positive impact on costs for the main stage.

If the retailer approach develops over the coming weeks, then this could also be included in the pilot testing.

4.4 Sample sizes

Below we set out two options and their strengths and weaknesses, stakeholders views and then our recommendations.

Options

Small companies are concerned that not enough sample can be recruited, particularly non-household customers. However, large companies would like to get sub-company results rather than just whole-company results; for example: by region (within company); hard-to-reach customers; future customers; low-income customers; vulnerable customers.

One possibility is to offer companies different sample size options, e.g.:

- 1,000 households and 500 non-households, including all above segmentations as required by the company, in addition to company-wide estimates.
- 500 households and 250 non-households, solely to obtain a company-wide set of estimates with no segmentations

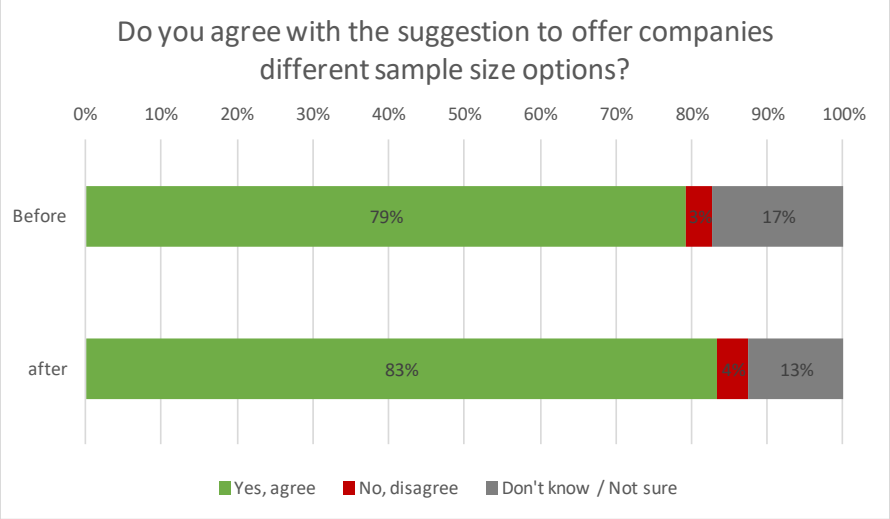
Another possibility is to include wastewater and water issues together for every customer, including those where services are supplied by separate companies. This would have implications for sample sizes with some companies having very large sample sizes in total.

Stakeholder views

The figures below show the views of workshop participants about sample sizes and the integration of wastewater and water issues.

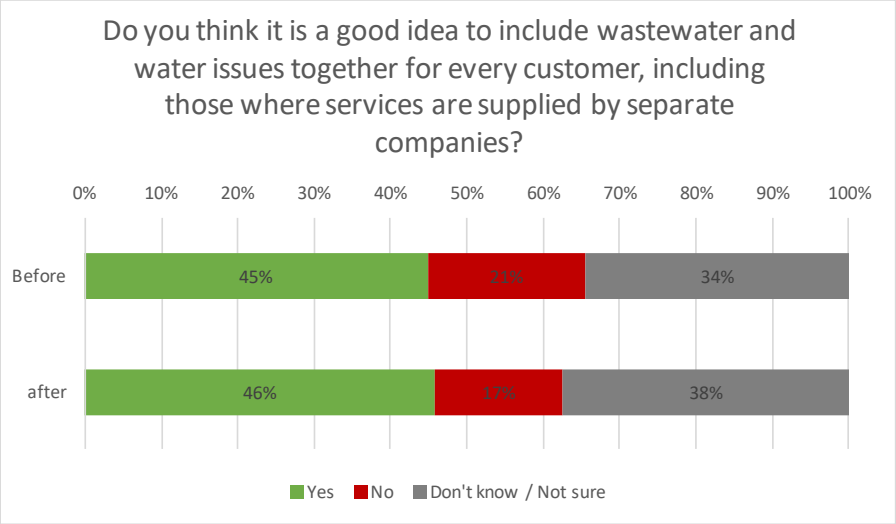
The vast majority of participants (79% before and 83% after the discussions,) agree with the suggestion to offer companies different sample size options.

Figure 23: Stakeholder views about sample sizes



45-46% of participants agreed with that it is a good idea to include wastewater and water issues together for every customer. Fewer participants disagreed (21% and 17%, before and after the discussions, respectively).

Figure 24: Stakeholder views about the integration of wastewater and water issues



During the discussions, there was consensus on offering different sample sizes to different companies. Participants agreed that the appropriate sample size depend on size of the company.

Regarding including water and sewerage issues, the main point of discussion was that if the customers' bottom priorities are all sewerage-related, it is not clear how the analysis could estimate the relative water-related priorities.

Recommendation

As discussed in Section 3, we are recommending that wastewater and water issues are included together for every customer.

We further recommend that there should be a minimum of 500 households in each water supply area, and in each wastewater supply area, for all companies except for Hafren Dyfrdwy for whom a smaller sample size of 350 is recommended for reasons of proportionality. These recommendations are based on the judgement that a minimum sample of around 150 participants is needed per segment estimate, and that a sample of 500 would thereby result in an appropriate level of precision on 3-way segmentation estimates. For example, age and socio-economic grade could be split into three categories with separate estimates obtained for each within the 500 sample size.

To the extent that companies wish to pay for a larger sample size, we recommend that they should be free to do so within reasonable limits based on the practicality of timescales. An across-the board boost could be purchased or, alternatively, larger sample sizes could be tailored to boost customer segments based on postcode, in the case of the PAF-based sampling approach, or to boost certain socio-demographic quotas in the case

of the online panel approach. Companies should have the flexibility to boost sample sizes in a manner best suited to how they wish to research their customer base.

In order to allocate sample, we have assembled a GIS dataset of water and wastewater supply area boundaries (source: Ofwat), overlaid this with Census population data, and derived the population sizes within each water-waste area. This included some very small cells, which we believe to be due to the inexact nature of the boundary shapefiles. It also included small suppliers such as Albion Water, Icosa, and Veolia Water Projects whose customers are not to be included in the Collaborative ODI research. After having removed these cells from the data, we were left with 31 water-wastewater supplier combinations.

Using this dataset, we allocated the sample for the water supply area proportionally in line with population between wastewater providers within the same area. For example, the SES Water sample is split 93:7 between Thames and Southern wastewater supply areas due to the balance of the respective populations.

In most cases, this means there are naturally at least 500 wastewater interviews per wastewater company. However, this is not the case for three companies: South West Water, Northumbrian Water and Hafren Dyfrdwy. In the former two cases, a boost has been applied to dual-service customers of these companies to bring the wastewater sample up to the minimum of 500.

In the case of Hafren Dyfrdwy, however, a minimum sample size of 500 wastewater customers was considered to be disproportionately large due to the very small size of the wastewater customer base. In this case, a boost was applied to bring the sample size up to 150, the minimum considered to be needed for a robust single-segment estimate. This will preclude segmentation of wastewater service valuations within the Hafren Dyfrdwy supply area, but will allow a company level estimate to be obtained.

For non-households, we have applied a global multiplier of 0.4 to all sample cell sizes in order for the total sample to be at least 200 for each water and wastewater supply area. Whilst, based on statistical considerations alone, the sample size for non-households should be at least as large as that for households, a smaller sample size has been recommended due to the fact that non-household interviews are much more costly per interview than household interviews. Accordingly, on an efficiency-per-pound basis, a smaller sample size for non-households than for households is appropriate.

In most cases, the recommended sample sizes for non-households will not allow for any segmentation of values for this customer group. It will allow robust estimates at a company level, however, for all companies except for Hafren Dyfrdwy.

For Hafren Dyfrdwy, the achievable non-household sample size for the wastewater service area is likely to be too small to estimate reliable values at even a company level. In this case, we recommend including additional observations outside the Hafren Dyfrdwy wastewater supply area, e.g., including those supplied by water by Hafren Dyfrdwy but wastewater by Welsh Water and/or combining household and non-household sample together. By so doing, an improved estimate should be possible. This analysis will need

to be developed carefully and transparently by the party responsible for this part of the Collaborative ODI research.

Table 8 contains the resulting sample plan, covering households and non-households. In this table, the company total is less than the sum of water and wastewater interviews due to the presence of dual-supply customers.

Table 8: Recommended base household and non-household sample sizes, by company

	Households			Non-households		
	Water	Waste	Company total	Water	Waste	Company total
Affinity Water	500		500	200		200
Anglian Water	507	767	804	203	307	509
Bristol Water	500		500	200		200
DÅµr Cymru	500	700	700	200	280	480
Hafren Dyfrdwy	350	150	350	140	60	200
Northumbrian Water	693	500	709	277	200	477
Portsmouth Water	500		500	200		200
SES Water	500		500	200		200
Severn Trent Water	500	920	937	200	368	568
South East Water	500		500	200		200
South Staffordshire Water	500		500	200		200
South West Water	605	500	605	242	200	442
Southern Water	500	1357	1380	200	543	743
Thames Water	500	1670	1670	200	668	868
United Utilities	500	500	500	200	200	400
Wessex Water	500	1089	1089	200	436	636
Yorkshire Water	500	502	517	200	201	401
Total	**8654**	**8654**	**12262**	**3462**	**3462**	**6924**

5 Analysis and outcomes

This section includes the following parts:

- Weighting of data
- Analysis and outcome expectations by choice exercise
- Linking the outcomes from the choice exercises
- Further research required
- Links to company research

5.1 Weighting of data

Upon receipt of the final datasets, weights should be generated to correct for non-response bias as well as for the uneven geographic distribution of the sample. In the latter case, this uneven distribution is caused by the sample plan (in Table 8), which draws an unequal proportion of each water-waste supplier combination for the purposes of ensuring that there are the correct numbers in total for each water company. Weights can, and should, be applied to correct for this sample distortion. The weights themselves will depend on the actual sample numbers in each cell of Table 8, which itself will be influenced by whether, and how much, companies decide to boost their sample numbers.

With respect to non-response bias, for households, this will involve comparisons of demographic variables including age, gender and socio-economic status against Census comparators for each company region. For non-households, weights should be derived based on a comparison of sample frequencies against population frequencies with respect to the size of the customer, measured by number of employees, against BEIS Business Population Estimates data for the closest corresponding region.

Once weights have been created, they should be used for all subsequent analysis.

5.2 Analysis and outcome expectations by choice exercise

In the following, we provide a high-level overview of the core type of analysis that is expected to be undertaken to obtain the required valuations for ODI rate setting. The econometric analysis stage is expected to be competitively tendered and so full details of the analysis that will be undertaken will follow as an outcome of this process.

Impact-based choice exercise

The data obtained from the proposed impact-based exercise reveal which of the two service issues shown on each choice occasion would have the most impact on the

participant. Taking all choice situations for all participants together, this results in a rich dataset on the relative impact of all the service issues.

A logit regression framework is appropriate as the starting point for the analysis. Within this framework, given service issues A and B, the probability that A is chosen as the most impactful is given by $\exp(V_A)/(\exp(V_A)+\exp(V_B))$. Here V_A and V_B are coefficients that are estimated by the maximum likelihood. (See, e.g. Train, 2003).

Such a model allows as many different service issue variables as are included in the exercise except that one must be omitted to serve as the base category. Let V_Z be the omitted service issue. By construction, Vz is then equal to zero, and $\exp(V_Z)=1$.

The values $\exp(V_A)$, $\exp(V_B)$, etc., are interpreted as the impacts of A, B, etc. relative to the impact of Z. This interpretation is consistent with the idea that the impact of A relative to B is equal to the probability of choosing A from the pair {A,B} divided by the probability of choosing B from the same pair.

An 'impact index' is then created by scaling the relative impacts $\exp(V_A)$, $\exp(V_B)$, etc., to sum to 1. This scaling is applied to avoid the choice of omitted service issue influencing the scale of the entire set of results and confusing any comparisons between segments of the relative impacts of different service issues.

The outcome from this basic analysis is an impact index bounded by [0,1].

Thus far, we have not yet imposed any structure to allow for different preferences for different people. This structure can be applied in a number of ways:

- Applying robust/clustered standard errors – this does not change the coefficients from the model but does adjust the estimated standard errors to allow for the fact that repeated choices are observed for each survey participant.
- Estimating separate models for different sample segments.
- Estimating models with interaction terms to allow for, and test, differences in coefficients for different segments.
- Estimating mixed logit models, which allow for unobserved preference heterogeneity by estimating a distribution of coefficients for the sample, where the type of distribution (e.g. normal, lognormal) is specified by the analyst.

The last of these approaches is particularly significant as it allows for individual-level coefficients, and hence impact indices, to be inferred from their choices conditional on the distribution of coefficients estimated for the sample as a whole. (See Train, 2003, ch.11.) Given individual-level coefficients it is possible to derive, and test for differences between, any segmentation of the sample within a single model.

Validity appraisal

The validity of the outputs of the quantitative analysis should be assessed along the usual two broad dimensions; content and construct validity.

Content validity judgements take into account the entirety of the study with the key test being that valid values are revealed by participants in the stated preference survey.

Examples of content validity analysis would include examination of responses to follow-up questions (following the valuation section). These should be analysed to identify cases where answers are invalid, for example due to participants misunderstanding the valuation task. Such cases will be identified and highlighted, and the option taken to remove them from the econometric analysis.

Construct validity assessments take into account the extent to which the output conforms with prior expectations. Examples of construct validity testing would include analysing the internal consistency of response data with expectations. For example, the impact of river pollution should be greater for those who use rivers for recreation, and likewise for coastal bathing water quality. Additionally, the relative impacts of different service issues should also conform to expectation; e.g., longer interruptions should have a greater impact than shorter interruptions

Compensation-based choice exercise

The compensation-based choice exercise will yield data similar to that derived from a double-bounded contingent valuation exercise. Taking both questions for a service issue together, the data imply an interval for the required compensation to accept the service issue in question. There are four possibilities, depending on the combination of answers:

- {*Interruption plus compensation; Interruption plus compensation*} implies a value bounded by £0 and the compensation amount offered in the second question.
- {*Interruption plus compensation; No interruption*} implies a value bounded by the compensation amount offered in the second question and the compensation amount offered in the first question.
- {*No interruption; Interruption plus compensation*} implies a value bounded by the compensation amount offered in the first question and the compensation amount offered in the second question
- {*No interruption; No interruption*} implies a value with a lower bound of the compensation amount offered in the second question, and no upper bound.

This data suggests an interval modelling strategy, as was originally proposed by Hanemann et al (1991). By so doing, the approach yields substantially more statistically efficient estimates in comparison to single-bounded dichotomous choice contingent valuation data, where there is only one response per person. Unfortunately, evidence has shown that the first and second choices are not independent of one another. Hence, more sophisticated methodologies have been proposed to handle the data, including bivariate probit (Cameron and Quiggin, 1994), random effects models (Alberini et al., 1997) and Bayesian updating models (McLeod and Bergland, 1999).

There are thus a range of modelling approaches suitable for analysis of such data, and it will be for the analyst completing the econometric modelling stage of the Collaborative ODI research to select a preferred method from this suite based on an exploratory analysis of the data.

Regardless of which modelling methodology is ultimately adopted, the outcome from this part of the analysis will include required compensation-based valuations of each of the service issues for each sample segment analysed.

Validity appraisal

As for the impact-based exercise, the validity of the outputs of the quantitative analysis should be assessed along the dimensions of content and construct validity.

Content validity analysis should include examination of responses to follow-up questions (following the valuation section). These should be analysed to identify cases where answers are invalid, and the option taken to remove them from the econometric analysis.

Construct validity analysis should check that required compensation correlated with income in the expected direction. It could also check that correlations are as expected between the impact weights derived for supply interruptions and external sewer flooding and the required compensation values derived for these service measures.

5.3 Derivation of PC valuations

A number of steps are needed following the analysis of impact-based and compensation-based exercise responses:

- First, one needs to derive values for each water-wastewater company combination, for each customer type.
- Then, these values need to be weighted to represent the values of the service issues at a company level.
- Finally, the service issue values need to be converted to values per unit of change of the original PC measures for each company.

These steps are explained in the following.

Linking the outcomes from the choice exercises

Taking one water-wastewater company combination, and one customer type {HH, NHH}, the method for calculating the value per avoided service issue of each kind is straightforward. From the compensation-based exercise, there will be estimates of mean and median values per avoided short supply interruption, and per avoided external sewer flooding incident. Say these are £100 and £500 respectively for the segment in question.

From the impact-based exercise analysis, suppose we have an estimated impact index for the segment in question that has impact scores of 0.5, 3 and 30 for a short supply interruption, an external sewer flooding incident and an internal sewer flooding incident respectively.

Combining these results gives the values shown Table 9. Here, the lower bound values are derived by pivoting off the estimated £500 compensation value for avoiding external

sewer flooding; the upper bound values are derived by pivoting off the estimated £100 compensation value for avoiding a short supply interruption. In both cases, the values of the remaining service issues are derived such that the values are proportional to the impact index derived from analysis of the impact-based exercise.

Table 9: Worked example combining outcomes from impact-based and compensation-based exercises (households)

Service issue	Impact index	Lower bound value	Upper bound value
Short supply interruption	0.5	£83	**£100**
External sewer flooding	3	**£500**	£600
Internal sewer flooding	30	£5,000	£6,000

Lower bound value based on compensation required to avoid External sewer flooding
Upper bound value based on compensation required to avoid a Short supply interruption

In the above table, only one additional service issue is shown, Internal sewer flooding. The extension to the full set of service issues included in the survey is trivial.

In the case of non-households, as discussed in Section 3.5, values would be derived as percentages/multiples of the annual bill rather than as monetary amounts. These would need to be converted to monetary values using data on the average bill for non-households.

Generating company-level estimates

Taking as given a set of household estimates like that shown in Table 9 for each water-wastewater company combination, and a comparable set of estimates for non-households but denominated in percentages/multiples of the annual bill, the next task is to aggregate to company level.

This will need to be done separately for water and wastewater service issues. Table 10 below provides a worked example of how wastewater values would be combined across two areas. Area 1 in this example could be imagined to be dual-supply households, while Area 2 includes households that are supplied wastewater services only by the company. The combined household value in this case is simply the weighted average of the two.

Table 10: Worked example combining household values across supply areas

Service issue	Area 1 value (900k households)	Area 2 value (100k households)	Combined value
External sewer flooding	£500	£700	£520
Internal sewer flooding	£5,000	£8,000	£5,300

Combined value derived as a weighted average of Area 1 and Area 2 values.

Table 11 presents a similar worked example combining household and non-household values. Here, there are many fewer non-households (10k vs 1m), but non-households have a substantially higher value per avoided service issue. Nonetheless, in this case, the combined value is close to the household value.

Table 11: Worked example combining household and non-household values

Service issue	HH value (1m households)	NHH value (10k non-households)	Combined value
External sewer flooding	£500	£2,500	£520
Internal sewer flooding	£5,000	£12,000	£5,069

Combined value derived as a weighted average of HH and NHH values.

A similar approach would be applied to calculate the values for water service issues, but the areas, and the numbers of households and non-households in each, would be different in this case.

Converting service issue valuations to PC valuations

The final step of the core analysis would be to translate values per service issue into values per marginal unit change in the original common PC definitions. This mapping will make use of the relationships set out in Section 3.4, and the values assigned to the parameters therein.

In some cases, the mappings are straightforward. For example, in the case of supply interruptions, suppose the average company value per avoided 6-hour unexpected supply interruption was £100. In this case, the value per minute, aggregated over the whole company, would be 100*N/360 where N is the number of customers in total.

In other cases, the mappings are more complex, as set out in Section 3.4. These mappings will all need to be fully determined in order to use the values derived from the Collaborative ODI research to derive ODI rates.

5.4 Further research required

The methodology set out thus far is sufficient to derive ODI rates based on a robust measure of marginal value, subject to agreement of a robust set of mappings from service issues to PCs. However, we also recommend that an additional piece of research is undertaken to measure customer preferences with regard to the overall relationship between bills and service levels. This piece of research would serve to set global limits on the degree to which bills could increase during the forthcoming price control period. This is expanded on below.

Setting limits on ODI-driven bill increases

In contrast to much of the previous valuation research in the water sector, the approach recommended here does not include any package scaling of valuations – where valuations are scaled to be consistent with willingness to pay for a broad-ranging package of service improvements. This is because the approach to package valuation was considered to be unreliable due to its complexity. (See Section 3.2 for our detailed review of the strengths and weaknesses.)

Accordingly, it is possible within the recommended approach that, in the absence of any limits set on ODIs, that bills could increase more than customers would wish them to overall.

To address this issue, we recommend that a separate survey is undertaken, close to submission of the final business plan, and possibly as part of the acceptability testing of the plan. This survey would include a question designed to measure customers' broad preference between alternative bill-service profiles.

Figure 25 shows an example of the kind of question we have in mind. This example was drawn from research conducted by ICS on behalf of Anglian Water.

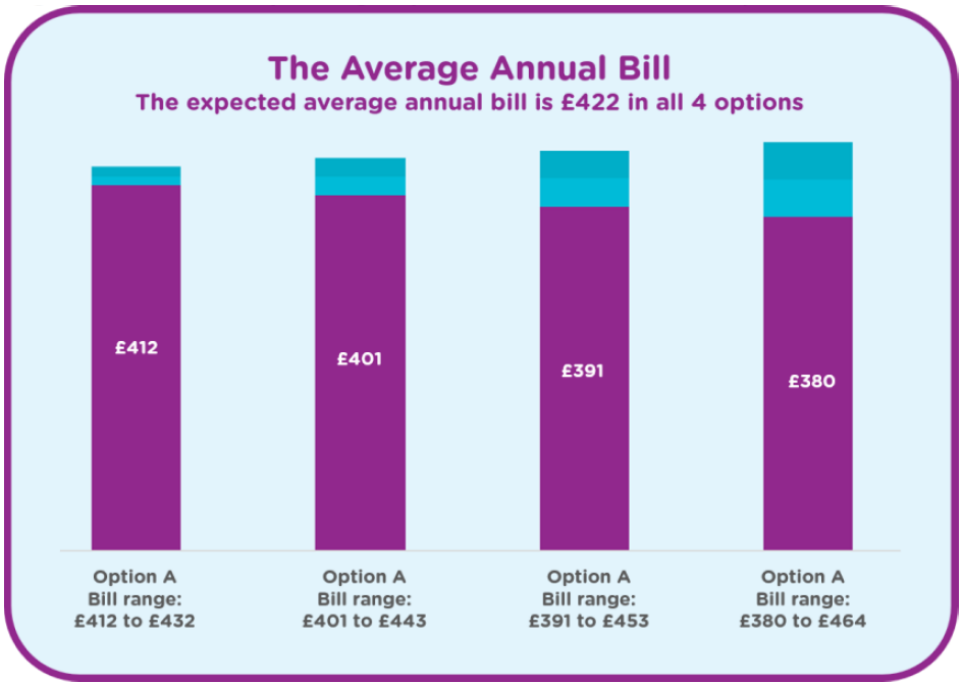
Following initial introductory material explaining the ideas behind ODIs, the survey would ask customers to rank the four options in order of preference. Each option has the same expected average annual bill but the options differ in how much bills could potentially vary over the price control period in line with service levels. Option D has the greatest variability of the bill while Option A has the least variability.

This question involves customers making a high level trade-off between:

- the benefits of increased incentives on companies to deliver better service, and
- the benefits of a stable bill.

These responses would then provide a customer preference justification for limiting the overall range of ODIs.

Figure 25: Example of approach to test preferred ODI range



Source: ICS (2018) for Anglian Water.

Importantly, the intention behind this exercise would not be to provide a scaling factor for the ODI rates derived from the Collaborative ODI research; nor would the approach

impose any caps or collars against individual PCs. The sole purpose of the approach would be to limit the overall bill increase that customers could potentially see as a consequence of the price control determination. Consequently, it would be appropriate to include this question alongside acceptability and affordability testing of the overall business plan.

5.5 Links to company research

The outcome from the Collaborative ODI research will be a set of valuations for a suite of common PCs. Ofwat has already strongly encouraged companies not to submit valuation evidence with a view to challenging ODI rates (Ofwat, 2021b), but expects that companies will potentially undertake further valuation research to support enhancement cases and/or bespoke ODI rates. In the following, we set out how companies' own research might best link into the outputs from the Collaborative ODI research to support these areas. 'While this goes beyond the scope of the ODI rates research addressed in detail in the rest of this report, we are here offering our views.

Enhancement cases

For PR24, companies are expected to develop enhancement cases via both long-term (25-year) and short-term (5-year) plans (Ofwat, 2021e), as well as develop ODIs that are to apply within the price control period. In this context, an important first question to address is whether there are differences in the types of values that should be obtained and, if so, how they might be reconciled. There is then a question as to what type of additional evidence might be needed and how they might this be generated in a manner consistent with the approach taken in the Collaborative ODI research.

Valuations for 25-year and 5-year CBA vs for ODI rates

In theory, the context for valuation – whether for long-term planning, short-term planning, or for setting ODI rates, should only matter to the extent that values are expected to grow or decline over time. Values should otherwise relate solely to the change in service level not to the planning context. If one expects values to change over time, which ought to be the case due to growth in incomes over time, the correct form of representation when applying the values should be as a time-series rather than as time-independent unit valuations per avoided service failure, or per unit of a PC measure.

The following worked example illustrates the idea here. Let Company A serve a population of 1 million bill-paying customers. Currently, there are 1,000 internal sewer flooding incidents per year. The company plans to reduce this number to zero over the course of 25 years, at an even rate. To pay for this, the company plans to increase bills by £1 per customer over the same period, so that bills increase in line with service.

In Table 12, two alternative versions of the cost-benefit analysis are shown:

- Case A) has a static valuation, in real terms, of £20k per avoided internal sewer flooding incident;
- Case B) has a value that increases, in real terms, at a rate of 2% per year.

When discounting over time using the Green Book social rate of time preference (HM Treasury, 2020), Case A returns a net present value (NPV) of -£37.0 million, whereas Case B has a positive NPV of £18.5 million. This illustrates the potential significance of whether a flat or an increasing valuation is assumed over a 25 year period.

Table 12: Worked example of cost-benefit analysis

A) Static valuation per avoided sewer flooding incident

	Year				
	0	1	2	..	25
Internal sewer flooding incidents	1,000	960	920		0
Value per avoided incident (£k)	20	20	20		20
Benefit of reduction (£k)		800	1,600		20,000
Bill impact per customer		1	2		25
Discount factor	1	0.966	0.934		0.423
NPV (£x)	**-37,027**				

B) Increasing valuation per avoided sewer flooding incident

	Year				
	0	1	2	..	25
Internal sewer flooding incidents	1,000	960	920		0
Value per avoided incident (£k)	20	20.4	20.8		32.8
Benefit of reduction (£k)		1,061	2,164		42,656
Bill impact per customer		1	2		25
Discount factor	1	0.966	0.934		0.423
NPV (£x)	**18,515**				

Based on 1 million bill-paying customers.

In practice, we would not expect the valuations to be the same when derived using different contexts and methods. We would not expect, for example, the value obtained from the compensation-based exercise to be the same as that estimated from a discrete choice experiment-based approach focused on choices of service levels and bills, even in the absence of any anticipated change in valuations over time. This is because there are other factors in play that have a substantial impact on stated preference valuation estimates including, most notably, the fact that people are insensitive to the scope of service change shown, to a large extent, in the context of the sorts of service levels usually included in water and wastewater valuation surveys. This insensitivity impacts on valuation estimates to such an extent as to cast serious doubt on the validity and reliability of the measures obtained under this method. (See Box 1.)

Stated preference valuations are thus inherently sensitive to the way in which they are derived, and it is hence true that questions around long-term service and bill levels could be expected to result in different implied valuations than when questions are around a 5-year programme of service and bill level changes. It is also the case that these are likely to be different than the valuations that are derived from questions about required compensation, as recommended in the present study.

There are a number of different ways one could reflect on this divergence in valuations. Firstly, one could imagine using different values for different purposes, i.e.:

- Required compensation values for ODIs
- Long-term values for long-term plan
- Short-term values for short-term plan

This approach is not recommended, due to the fact that it creates a dynamic inconsistency. This is where companies have a long-term plan to do one thing, but a short-term incentive to do something different. In general, for short-term incentives to be aligned with a long-term plan, the same values, which may be time dependent, need to be used in both settings.

Hence, when considering how the value per avoided sewer flooding incident ought to be measured for the purposes of 5-year or 25-year plans, the key objective ought to be the estimation of:

- A) a current value; and
- B) how that value should be expected to evolve over time.

There are many ways in which value estimates can be obtained, each of which has strengths and weaknesses, as have been reviewed in Section 3.2. Whilst there is merit, in principle, to considering other approaches, this potentially comes at the cost of comparability across companies, which could favour use of a single common set of values to be used enhancement cases based around common PCs, as well as for the setting of ODI rates.

If this argument is accepted then, with respect to the long-term valuation of enhancement programmes based around common PCs, the only outstanding requirement for CBA beyond the values that will emerge from the Collaborative ODI research is an understanding of how values might be expected to evolve over time. As shown in the above worked example, the assumption made regarding this growth rate could be economically significant within the justification of a long-term plan.

In order to set the assumption, one option would be to engage customers. For example, customers could, in principle, be asked questions designed to elicit how they expect their own values to change over time.

It is doubtful, however, that a valid and reliable form of such a method could be developed. This is due to the fact that customers could be expected to find it difficult to know themselves what their preferences will be way into the future. As noted by Blue Marble, in their review of PR19 research for CCW:

'Consumers find it difficult to give informed responses in research that focusses on the future. (…) Consumers, when faced with a future scenario in business planning research, are often hesitant to commit to a point of view because they know that no one can anticipate the social, political, economic and environmental factors that will exist in 20-30 years' time. (…) Rooting research in consumers' current and historic experiences, and extrapolating from this where necessary, may be more valid in some instances.' Blue Marble (2020), p.31.

An alternative approach, which would appear to be in line with the Blue Marble recommendations, would be to extrapolate into the future based on one of two methods:

- Trends in valuations over time. (This is one for the future rather than for PR24 as it depends on the recommended compensation-based method being applied on multiple occasions over time.)
- Estimation of income elasticity of values via cross-section analysis of national Collaborative ODI research survey, and applying this elasticity to a GDP per capita forecast. For example, if the income elasticity of the required compensation is 0.5, i.e., if 2% higher income across the sample leads to 1% higher valuations, then, given a real GDP per capita forecast of 2% per year over 25 years, the unit values could be predicted to rise by 1% per year over the same period.

In our view, this approach would be preferable as a means of setting the growth rate of valuations over time to approaches based on using discrete choice experiments, or other methods, that require participants to imagine far into the future.

While only the current value and its growth rate are required for CBA, it would be good practice to also test high level choices of bill and service levels with customers as part of the long-term and short-term plan-testing process. This testing would not be intended to generate valuation estimates but, rather, to cap bill increases at levels that most customers say they would be willing to pay overall.

In summary, when thinking about the application of values in short-term versus long-term contexts:

- It is important that the same values are used for the same common PCs, rather than values derived from different stated preference methods. This is primarily to ensure that short-term incentives are aligned with a long-term plan.

- The values themselves are likely to be time-dependent and it is important that the growth factor over time in values is appropriately set.

- In addition to CBA, it is important to give customers choices between overall bill-service profiles rather than letting the profiles emerge directly from the CBA. This is in order to allow customers to exercise some influence over the broad direction of travel of their water and wastewater service.

Additional valuation evidence required

While the above has considered requirements with respect to common PCs, enhancement cases will often need to draw in evidence on a wider set of benefits, and environmental and social costs. Much of this work will be, or will have been, undertaken as part of the WRMP or DWMP processes. For example, investments may have wider recreation or local regeneration benefits; or they may have negative local disruption impacts, or specific local environmental impacts.

In general, the method proposed for the Collaborative ODI research does not impose any requirements on how additional evidence should be generated. There are no package effects, for example, whereby values obtained from new evidence need to be scaled to

be consistent. Instead, to the extent that values are validly and reliably measured, and do not double-count any benefits already accounted for via the common PC valuations, they can simply be added on to these values.

This principle is thus unrestrictive regarding the adding together of well-derived valuations for the purposes of completing a CBA. However, as part of the validity appraisal process, it would be good practice to draw comparisons between the valuations obtained from different sources, where possible, as a check that the relative values conform with expectation - for example, longer term supply interruptions should have a higher value than shorter interruptions.

As previously indicated, it is also important to test high level choices of bill and service levels with customers as part of the long-term and short-term plan-testing process. This testing would not be intended to generate valuation estimates, nor contribute to the CBA but, rather, to cap bill increases at levels that most customers say they would be willing to pay overall.

Bespoke ODIs

For PR24, there are expected to be fewer bespoke ODIs than at PR19 (Ofwat, 2021a). However, for the bespoke ODIs that are developed, a similar principle applies to the evidence needed as in the case of enhancement case evidence: to the extent that values are validly and reliably measured, and do not double-count any benefits already accounted for via the common PC valuations, they can simply used without any direct linkage to the common PC valuations.

In the case of some bespoke ODIs, however, it may be appropriate to adopt an impact-based approach to valuation in order to ensure consistency of valuations between bespoke and common ODIs. This would be particularly the case when bespoke ODI may be straightforwardly and reliably linked to customer impacts.

For example, suppose a company were to have an ODI based on the number of properties affected by persistent low pressure. In this case, it would be appropriate to research the relative impact of persistent low pressure in comparison to a short supply interruption and/or external sewer flooding, in order to allow valuations to be 'pivoted' off these values in the same way as for the common PC ODIs as described in 5.3 above.

In other cases, there could be merit in examining additional evidence to supplement evidence on relative impacts. For example, suppose a company were to have an ODI linked to local traffic disruption. The evidence needed to support this ODI would include a valid and reliable measure of the marginal impact of avoiding traffic disruption, which could be obtained by establishing the relative impact of local traffic disruption in comparison to a short supply interruption and/or external sewer flooding. This would allow valuations to be obtained in the same framework as for the common PC ODIs. However, alternative approaches are also available for this measure, including wellbeing-based valuations of avoided roadworks (see Fujiwara et al., 2021); and/or valuations based on the number of hours spent in traffic combined with DfT estimates of the value of travel time (DfT, 2021). A better evidence base would draw on multiple sources of evidence provided these show good standards of validity and reliability.

6 Next steps

6.1 Workplan

The next stage of the study involves testing and refining the methodology. The following table sets out the work plan agreed for this next phase of the study.

Table 13: Project work plan (Stage 2)

Ofwat to provide latest information on likely performance commitments	January 2022	Ofwat
Share methodology and draft research materials	Late January	Accent-PJM
Feedback on methodology and draft research materials	Early February	Ofwat, CCW, water companies
Companies and regulators to provide information required by the methodology regarding service issue definitions	Early February	Companies, DWI, EA, Ofwat
Test the research materials with household and non-household customers	February -April	Accent-PJM
Industry workshop on final research materials	April	Accent-PJM, Ofwat, CCW, water companies, others
Feedback on research materials	April	Ofwat, CCW, water companies
Research materials finalised for each water and wastewater company	End of April	Accent-PJM

As shown in the table, input from companies is sought at a number of stages:

- To review and comment on the recommendations and draft research materials in Early February
- To provide information required by the methodology regarding service issue definitions in Early February.
- At an industry workshop, to be held via Teams, in April 2022.

6.2 Final Stage 2 outputs

The final output from the study will include a complete set of well-tested research materials, with accompanying experimental designs and supporting documentation, for obtaining the evidence needed to support the setting of ODI rates for common PCs at PR24.

References

Accent (2017) Measures of success: Quantitative findings (version 3). Report to Dwr Cymru Welsh Water. November 2017.

Accent-PJM Economics (2014) Comparative Review of Willingness to Pay Results. Report for a club of UK water companies.

Accent-PJM Economics (2017) Dŵr Cymru Welsh Water PR19 Willingness to Pay Research. Report for Dŵr Cymru Welsh Water.

Accent-PJM Economics (2018a) Exploration of Supply Outage Compensation Levels. Report for Affinity Water. June 2018.

Accent-PJM Economics (2018b) Comparative Review of PR19 WTP Results. Report for a club of UK water companies.

Accent-PJM Economics (2018c) Shopping Basket Research. Report for Southern Water. January 2018.

Aecom-DJS (2018) PR19 Understanding Customer Values: Work Package 1 – First Round Stated Preference. Report for Yorkshire Water.

Aecom (2017) PR19 Understanding Customer Values: Work Package 5 – Behavioural Experiment. Report for Yorkshire Water.

Alberini, A., Kanninen, B. and Carson, R. T. (1997) Modelling response incentive effects in dichotomous choice contingent valuation data. Land Economics, 73, 309–324.

Ariely, D., Loewenstein, G. and Prelec, D. (2003) "Coherent arbitrariness": Stable demand curves without stable preferences. Quarterly Journal of Economics, 118, 73-105.

Bateman, I. J. and Willis, K. G. (eds.) (1999) Valuing Environmental Preferences. Oxford, UK.

Bateman, I., Carson, R., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Özdemiroğlu, E., Pearce, D., Sugden, R. and Swanson, J. (2002) Economic Valuation with Stated Preference Techniques: A Manual. Edward Elgar, Cheltenham, UK.

Blue Marble (2020) Engaging water customers for better consumer and business outcomes. Report for CCW.

BoxClever (2018) Acceptability testing for PR19 stage 1. Report for United Utilities.

Boyle, K., Bishop, R. and Welsh, M. (1985) Starting Point Bias in Contingent Valuation Bidding Games. Land Economics, 61(2),188-194.

Burrows, J., Newman, R., Genser, J. and Plewes, J. (2017) Do contingent valuation estimates of willingness to pay for non-use environmental goods pass the scope test with adequacy? A review of the evidence from empirical studies in the literature. In McFadden, D. and Train, K. (eds) (2017) Contingent Valuation of Environmental Goods: A Comprehensive Critique. Edward Elgar, Cheltenham, UK.

Bergstrom, T. C. (1982) When is a man's life worth more than his human wealth? In M. W. Jones-Lee (ed.) (1982) The value of life and safety. Amsterdam, North-Holland.

Cameron, T. and Quiggin, J. (1994) Estimation using contingent valuation data from a 'dichotomous choice with follow-up' questionnaire. Journal of Environmental Economics and Management, 27, 218–234.

Carson, R. T. (2012) Contingent valuation: A practical alternative when prices aren't available. Journal of Economic Perspectives, 26(4), 27–42.

Carson, R. T. and Groves, T. (2007) Incentive and informational properties of preference questions. Environmental and Resource Economics, 37, 181-210.

Cascade (2011) The Use of Revealed Customer Behaviour in Future Price Limits. Report for Ofwat.

Chalak, A. and Metcalfe, P. (2021) Valuing water and wastewater service improvements via impact-weighted numbers of service failures. Journal of Environmental Economics and Policy. Forthcoming.

Day, B., Bateman, I., Carson, R., Dupont, D., Louviere, J., Morimoto, S., Scarpa, R. and Wang, P. (2012) Ordering effects and choice set awareness in repeat-response stated preference studies. Journal of Environmental Economics and Management, 63, 73-91.

DWI (2006) Annual provision of information on consumer contacts. Information Letter 1/2006. 6 Jan 2006.

DWI (2018a) DWI Compliance Risk Index (CRI). Aug 2018.

DWI (2018b) DWI Event Risk Index (ERI). Aug 2018.

Freeman, A. M. (2003) The Measurement of Environmental and Resource Values, 2nd Ed. RFF Press, Washington DC, USA.

Fujiwara, D., Houston, R., Keohane, K., Maxwell, C. and van Emmerik, I. (2021) Applying the wellbeing valuation method to value the costs of roadworks and flooding, Journal of Environmental Economics and Policy, DOI: 10.1080/21606544.2021.1938688.

Green, D. Jacowitz, K., Kahneman, D. and McFadden, D. (1998) Referendum contingent valuation, anchoring, and willingness to pay for public goods. Resource and Energy Economics, 20, 85-116.

Hanemann, W. M., Loomis, J. and Kanninen, B. (1991) Statistical efficiency of double bounded dichotomous choice contingent valuation. American Journal of Agricultural Economics, 73, 1255–1263.

Hausman, J. (2012) Contingent valuation: From dubious to hopeless. Journal of Economic Perspectives, 26(4), 43–56.

Herriges, J. and Shogren, J. (1996) Starting Point Bias in Dichotomous Choice Valuation with Follow-Up Questioning. Journal of Environmental Economics and Management, 30, 112-131.

HM Treasury (2020) The Green Book: Central Government Guidance on Appraisal and Evaluation.

ICS (2018) Outcome Delivery Incentive Research. Report for Anglian Water.

ICS-Eftec (2018) PR19 Main Stage Willingness to Pay Study. Report for Anglian Water.

Kopp, R. (1992) Why existence value should be used in cost-benefit analysis. Journal of Policy Analysis and Management, 11(1), 123-130.

Lanz, B, Provins, A., Bateman, I., Scarpa, R., Willis, K., and Ozdemiroglu, E. (2010). Investigating willingness to pay – willingness to accept asymmetry in choice experiments, in S. Hess and A. Daly (eds.) Choice Modelling: the state-of-the-art and the state-of-practice: proceedings from the inaugural International Choice Modelling Conference. Emerald Publishers, Bingley. pp. 517-542.

Lanz, B., and Provins, A. (2016) The Demand for Tap Water Quality: Survey Evidence on Water Hardness and Aesthetic Quality. Water Resources and Economics 16: 52–63.

Lopes, A. F. and Kipperberg, G. (2020) Diagnosing insensitivity to scope in contingent valuation. Environmental and Resource Economics, 77, 191-216.

McLeod, D. M. and Bergland, O. (1999) Willingness-to-pay estimates using the double-bounded dichotomous-choice contingent valuation format: a test for validity and precision in a Bayesian framework. Land Economics, 75, 115–125.

McConnell, K. E. (1997) Does Altruism Undermine Existence Value? Journal of Environmental Economics and Management, 32, 22-37.

McFadden, D. and Train, K. (eds) (2017) Contingent Valuation of Environmental Goods: A Comprehensive Critique. Edward Elgar, Cheltenham, UK.

Louviere, J., Flynn, T. and Marley, A. (2015) Best-worst scaling: Theory, methods and applications. Cambridge, UK.

Metcalfe, P. J. and Baker, W. (2011) Willingness to Pay to Avoid Drought Water Use Restrictions. Working paper. DOI:10.13140/RG.2.1.1449.3844.

Metcalfe, P. J., Baker, W., Andrews, K., Atkinson, G., Bateman, I., Butler, S., Carson, R., East, J., Gueron, Y., Sheldon, R. and Train, K. (2012), An assessment of the nonmarket benefits of the Water Framework Directive for households in England and Wales, Water Resour. Res., 48, W03526, doi:10.1029/2010WR009592.

Metcalfe, P. J. and Sen, A. (2021) Sensitivity to scope of water and wastewater service valuations: a meta-analysis of findings from water price reviews in Great Britain, Journal of Environmental Economics and Policy, DOI: 10.1080/21606544.2021.1984314.

NERA-Accent (2011) Carrying out willingness to pay research. Report for UK Water Industry Research.

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

Ofwat (2018) Reporting guidance – Sewer flooding.

Ofwat (2021a) PR24 and beyond: Creating tomorrow, together. May 2021.

Ofwat (2021b) PR24 and beyond position paper: Collaborative customer research for PR24. October 2021.

Ofwat (2021c) PR24 and beyond: Performance commitments for future price reviews, November 2021.

Ofwat (2021d) Common PCs for inclusion in the ODI rates research. Personal communication from Simon Compton. 9 December 2021.

Ofwat (2021e) PR24 and beyond: Long-term delivery strategies and common reference scenarios. November 2021.

Orr, S. (2007) Values, preferences, and the citizen-consumer distinction. Politics, Philosophy and Economics, 6(1), 107-130.

Ove Arup (2008) A framework for cost-benefit analysis in odour control projects. Report for UK Water Industry Research.

Pearce, D., Atkinson, G. and Mourato, S. (2006) Cost-benefit analysis and the environment: Recent developments. OECD, Paris, France.

Rose, J. M. and Bliemer, M. C. J. (2009) Constructing Efficient Stated Choice Experimental Designs, Transport Reviews, 29:5, 587-617, DOI:10.1080/01441640902827623

Rosenthal, D. and Nelson, R. (1992) Why existence value should not be used in cost-benefit analysis. Journal of Policy Analysis and Management, 11(1), 116-122.

Smith, V. K., and Moore, E. M. (2010) Behavioral Economics and Benefit Cost Analysis. Environmental & Resource Economics, 46: 217-34.

Train, K. (2003) Discrete choice methods with simulation. Cambridge, UK.

Turner, R. K. (1999) The place of economic values in environmental valuation. In Bateman, I. J. and Willis, K. G. (eds.) (1999) Valuing Environmental Preferences. Oxford, UK.

United Utilities (2018) PR19 Business Plan.

United Utilities (2021) Developing a national approach to customer research: Proposal for an approach.

Von Neumann, J. and Morgenstern, O. (1953) Theory of Games and Economic Behavior. 3rd Edition, Princeton University Press, Princeton.

Whitehead, J. (2002) Incentive Incompatibility and Starting-Point Bias in Iterative Valuation Questions. Land Economics, 78(2), 285-297.

Willis, K. and Sheldon, R. (2021) Research on customers' willingness-to-pay for service changes in UK water company price reviews 1994–2019. Journal of Environmental Economics and Policy. DOI: 10.1080/21606544.2021.1927850.

Yorkshire Water (2018) Our PR19 Plan. Appendix 5e: Understanding Customer Values_Stated Preference Report.

Appendix A – Mapping PCs to service issues

The proposed research methodology requires a set of customer-impacting service issues linked to each of the common PCs, and an agreed quantitative relationship between them. Section 3.4 discussed the selection and definition of service issues linked to each of the common PCs for which values are required. This appendix sets out the algebraic relationships between service issues and PC measures in general terms, focusing on two potentially problematic areas:

- PC mapping while avoiding double counting.
- Mapping experience of water quality issues to contacts about water quality.

PC mapping while avoiding double counting

In general terms, we envisage a customer value function, V, that is decreasing in the likelihood of each of a number of different service issues impacting on their household or local area. In a customer survey, the values of avoiding each type of service issue can be measured using the proposed impact-based method of valuation.

Let X_1 and X_2 be two types of service issue that customers would like to avoid, let the probabilities of these occurring be r_1 and r_2 respectively, and let the values per incident of each type be a and b respectively. We then have the following customer valuation function:

(1) $V = ar_1 + br_2$

Next, we posit two PC measures, M_1 and M_2, that both impact on r_1 and r_2, but via distinct mappings, as shown below.

(2) $M_1 = w_{11}r_1 + w_{12}r_2$

(3) $M_2 = w_{21}r_1 + w_{22}r_2$

Here, we might imagine, for example, that M_1 is the Compliance Risk Index (CRI); M_2 is the Event Risk Index (ERI); X_1 is the receipt of a Boil water notice and X_2 is the receipt of a Do not drink notice. In this case the terms w_{11} and w_{12} represent the impact of a 1-unit change in the CRI on the likelihoods of a Boil water notice and a Do not drink notice respectively. Likewise, the terms w_{21} and w_{22} represent the impact of a 1-unit change in the ERI on the likelihoods of a Boil water notice and a Do not drink notice respectively.

Equation (4) expresses how customer value depends on changes in the PCs M_1 and M_2. It says that the change in the value, dV, is equal to the rate of change in value with respect

to M_1 holding M_2 constant, $\frac{\partial V}{\partial M_1}$, multiplied by the change in M_1, dM_1, plus the rate of change in value with respect to M_2 holding M_1 constant, $\frac{\partial V}{\partial M_2}$, multiplied by the change in M_2, dM_2.

(4) $$dV = \frac{\partial V}{\partial M_1} dM_1 + \frac{\partial V}{\partial M_2} dM_2$$

The partial derivatives $\frac{\partial V}{\partial M_1}$and $\frac{\partial V}{\partial M_2}$are the values that should be used to determine ODI rates for M_1 and M_2 respectively.

To calculate these values, we need to express V as a function of M_1 and M_2. To do so, we first derive r_1 and r_2 as functions of M_1 and M_2; then we combined these using Eq. (1).

From Eq. (2) and (3), we have:

(5) $$r_1 = \frac{M_1 - w_{12} r_2}{w_{11}}$$

(6) $$r_1 = \frac{M_2 - w_{22} r_2}{w_{21}}$$

(7) $$w_{21}(M_1 - w_{12} r_2) = w_{11}(M_2 - w_{22} r_2)$$

(8) $$r_2(w_{11} w_{22} - w_{21} w_{12}) = w_{11} M_2 - w_{21} M_1$$

(9) $$r_2 = \frac{w_{11} M_2 - w_{21} M_1}{w_{11} w_{22} - w_{21} w_{12}}$$

This gives r_2 as a function of M_1 and M_2. Similarly, we also have:

(10) $$r_2 = \frac{M_1 - w_{11} r_1}{w_{12}}$$

(11) $$r_2 = \frac{M_2 - w_{21} r_1}{w_{22}}$$

(12) $$w_{22}(M_1 - w_{11} r_1) = w_{12}(M_2 - w_{21} r_1)$$

(13) $$r_1(w_{12} w_{21} - w_{22} w_{11}) = w_{12} M_2 - w_{22} M_1$$

(14) $$r_1 = \frac{w_{12} M_2 - w_{22} M_1}{w_{12} w_{21} - w_{22} w_{11}}$$

Thus, r_1 and r_2 are given as functions of M_1 and M_2. Entering these into Eq.(1) yields:

(15) $$V = a \frac{w_{12} M_2 - w_{22} M_1}{w_{12} w_{21} - w_{22} w_{11}} + b \frac{w_{11} M_2 - w_{21} M_1}{w_{11} w_{22} - w_{21} w_{12}}$$

This can be rearranged to give:

(16) $$V = M_2 \left(\frac{a w_{12}}{w_{12} w_{21} - w_{22} w_{11}} + \frac{b w_{11}}{w_{11} w_{22} - w_{21} w_{12}} \right) - M_1 \left(\frac{a w_{22}}{w_{12} w_{21} - w_{22} w_{11}} + \frac{b w_{21}}{w_{11} w_{22} - w_{21} w_{12}} \right)$$

Taking the partial derivatives of V with respect to M_1 and M_2 gives:

$$(17)\quad \frac{\partial V}{\partial M_1} = -\left(\frac{aw_{22}}{w_{12}w_{21}-w_{22}w_{11}} + \frac{bw_{21}}{w_{11}w_{22}-w_{21}w_{12}}\right)$$

$$(18)\quad \frac{\partial V}{\partial M_1} = \frac{aw_{22}}{w_{22}w_{11}-w_{12}w_{21}} + \frac{bw_{21}}{w_{21}w_{12}-w_{11}w_{22}}$$

$$(19)\quad \frac{\partial V}{\partial M_2} = \frac{aw_{12}}{w_{12}w_{21}-w_{22}w_{11}} + \frac{bw_{11}}{w_{11}w_{22}-w_{21}w_{12}}$$

Hence, equations (18) and (19) can be used to derive the values to be assigned to PCs M_1 and M_2 as a function of:

- a and b, the value weights from the survey
- w_{11}, w_{12}, w_{21} and w_{22} the multiples of r_1 and r_2 that are associated with a 1 unit change in M_1 and M_2 respectively.

Accordingly, in order to value CRI and ERI in this example, it is necessary to agree the impacts of a 1-unit change in the CRI, and ERI, on the likelihoods of a Boil water notice and a Do not drink notice respectively. The survey will then be able focus solely on valuing avoided Boil water notices and Do not drink notices and the valuations for CRI and ERI will be derivable from the combined set of estimates.

There is an obvious question as to how generalisable this method is, in terms of how equations (1), (2) and (3) might feasibly vary in structure for different PCs and customer facing attributes. With respect to Eq. (1), there is no concern regarding generality as this format – with value equal to a probability weighted sum of service issue valuations – is a core feature of the proposed method for valuation.

With regards to equations (2) and (3), the approach holds for any case where a one-unit change in the PCs can be reasonably well expressed, for the purposes of valuation, as a weighted sum of probability changes with respect to any number of customer-facing measures.

It thus holds, via a simple extension to the algebra, if one includes three, four, or more different types of service issue as consequent to changes in the CRI or ERI, as long as one can specify the impact on the probability of each of these service issues occurring per 1-unit change in the PC itself.

Mapping experience of water quality issues to contacts about water quality

Let r_1 and r_2 be the risks of warned and unwarned water quality issues respectively, and let a and b be their associated unit valuations, both negative and with $|a|<|b|$. Furthermore, let $r_1=pr$ and $r_2=(1-p)r$ where p is the proportion of total water quality issues that are warned and r is the overall risk of experiencing either a warned or an unwarned water quality issue.

Then, value can be expressed as follows:

(20) $V = r\big(ap + b(1-p)\big)$

Next, let *M* be the number of water quality contacts. As shown below, this is structured as the weighted sum of the risks of warned and unwarned water quality issues, where the weights w_1 and w_2 are equal to the likelihoods of customers contacting the company upon experiencing each type of issue, multiplied by the number of customers. Here, both w_1 and w_2 are positive with $w_1<w_2$ – the likelihood of contacting the company is lower following a warning than when unwarned.

(21) $M = r\big(w_1 p + w_2(1-p)\big)$

The appropriate valuation to attach to *M* is given by *dV/dM*.

(22) $\frac{dV}{dM} = \frac{dV}{dr}\frac{dr}{dM} = \frac{\big(ap+b(1-p)\big)}{\big(w_1 p+w_2(1-p)\big)}$

This value depends on *p*. Given Eq. (22), an estimate is therefore needed for *p*, the proportion of total water quality issues that are warned, in addition to estimates of w_1 and w_2 in order to be able to assign a valuation to *dV/dM.*

Although *p* is under the control of the water company, it is still appropriate, in our view, to incorporate an estimate of *p* within the value function in Eq. (22). The alternative approach of measuring the value of the contact itself would involve an unsatisfactory disconnect from the way that customer value is actually driven.

When setting a value for *p*, it should be noted that value is strictly increasing in *p*. This is because:

(23) $\frac{dV}{dp} = r(a-b)$

This expression is positive because both *r* and *(a – b)* are positive; hence, value is highest when *p* = 1 (all warned) and lowest when *p* = 0 (none warned), holding *r* constant.

This suggests that companies should be encouraged to maximise *p*. However, this should not be at the expense of issuing an excessive number of warnings that turn out to be inaccurate: when the customer does not end up experiencing any water quality issue at their property despite receiving a warning from their water company. Although false warnings are not included in the model, one could expect these would have a negative impact on customers and so should therefore be avoided where possible.

Appendix B – Stakeholder views on survey design issues

At the 13 December 2021 industry workshop, a number of options were presented and discussed concerning various aspects of the survey design, including:

- Principle of translating PCs to customer facing measures
- Relative valuation of service areas
- Monetisation of relative values

This Appendix presents a summary of the views expressed, including polling results from the workshop.

Principle of translating PCs to customer facing measures

In the workshop, a distinction was drawn between two potential approaches to the specification of the research for valuing common PCs:

- Option 1: include PC measures directly
- Option 2: translate PC measures into customer-facing measures

Table 14 shows the examples of PC measures and corresponding customer-facing measures as presented in the workshop.

Table 14: Examples of PC measures and correspondent customer-facing measures

Examples of PC measure	Customer facing measure
Customer contacts about water discolouration, or taste and smell issues	Your tap water is discoloured for 1 day. It is safe to drink but you may choose not to. Your tap water has an unpleasant taste and smell for 1 day. It is safe to drink but you may choose not to.
Average supply interruption greater than three hours (minutes per property)	Your tap water supply stops working without warning and remains off for 12 hours.
Compliance risk index	You receive a notice from your water company instructing you to boil tap water before drinking, cooking or preparing food for 2 days. You receive a notice from your water company instructing you not to drink, cook or prepare food with your tap water for 2 days. Boiling will not purify the water.

PC measures were considered to have two advantages:

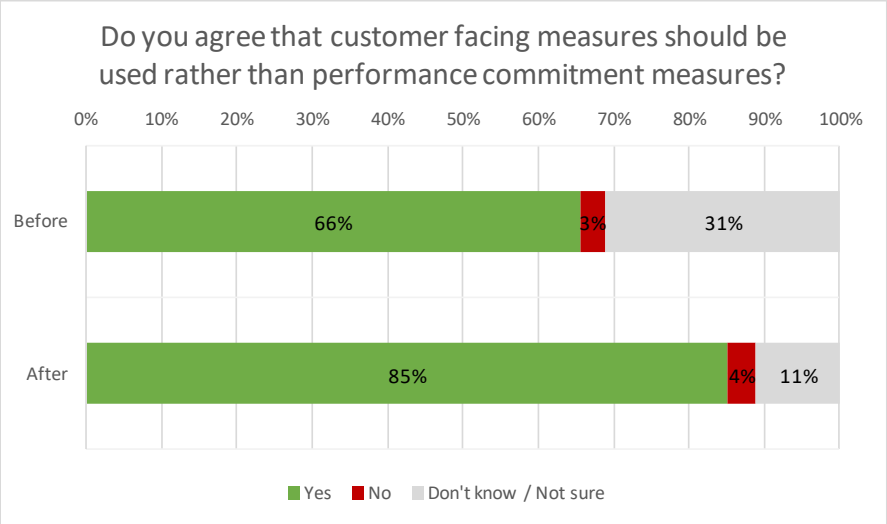
- They are directly aligned with the companies' price reviews
- They do not require assumptions by designers, analysts, or regulators

In contrast, customer-facing measures:

- are easier for customers to understand
- are more meaningful for customers
- do not require assumptions

The figure below synthesizes the views of the workshop participants regarding the use of customer facing measures. Before the workshop discussions, 66% agreed that customer-facing measures should be used and 3% disagreed. After the discussions, 85% agreed that customer-facing measures should be used and 4% disagreed.

Figure 26: Stakeholder views about customer-facing measures



In the breakout group discussions, while most participants agreed that customer-facing measures should be used, there was also some discussion on the challenges of this approach. Participants agreed that some measures are easier to translate than others.

The survey design needs to consider how the customer-facing measures can realistically be measured. Examples given by workshop participants included:

- Leakage - customers may not understand what it is
- Per-capita consumption reduction - customers do not usually value their consumption reduction or even somebody else's reduced consumption

A related issue was that, in some cases (e.g. water discolouration), the customer-facing measures are about the actual impacts on individual customers, but and in other cases (per capita consumption, supply interruptions, leakage), they are about the companies' overall performance across the supply area.

Another issue raised, in the context of per-capita consumption, was that, similarly to river water quality, it can be affected by actions in many sectors, so there was considered to be a need to disentangle the impact caused by the water sector.

In the case of compliance risk index, results were also said to be dependent on how survey questions are asked. It is different to derive values of not drinking poisonous water and values of not drinking slightly discoloured, but not dangerous, water. In reality, in most cases, customers are not at serious risk. The standards for water quality are placed very high but even companies not reaching the standards are delivering high performance, and the risk for customers is low. Small risk reductions may also not be perceived by customers.

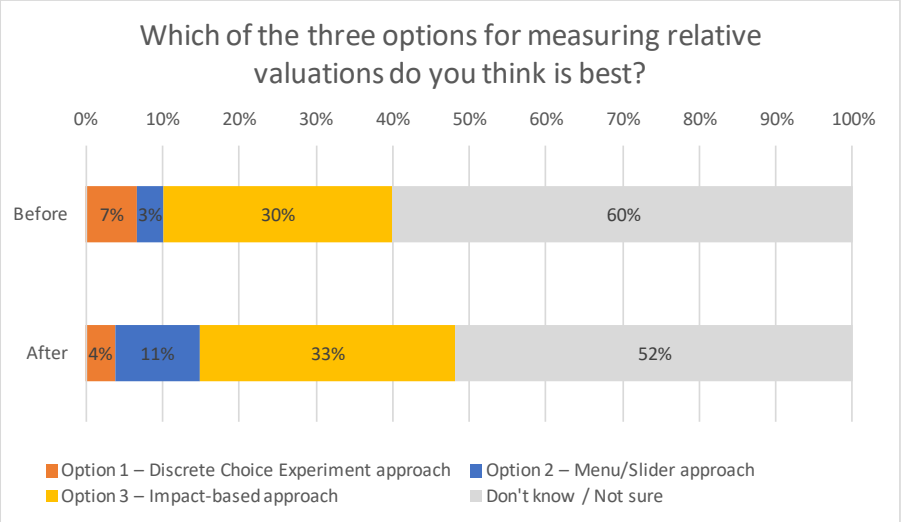
Compliance risk index can also cover a whole range of different impacts. The challenge is how to design a survey question that encompasses these impacts. If it focuses only on severe impacts, it can result on high estimated values. One solution could be showing a mid-level impact.

Despite these challenges, the overall view was that customer-facing measures are preferred because customers often do not understand the performance commitment concepts used by water companies. Customer-facing measures are preferable, even if it is difficult to translate well some of the measures. It is important, however, that customer-facing measures are aligned with the price review process and its requirements.

Approach for obtaining relative values

The figure below shows the views of workshop participants about stated preference methods. Option 3 (Impact-based approach) was the preferred option, out of the three options presented. However, the majority of participants (60%, before the discussion, 52% after the discussion) stated they did not know which option was the best.

Figure 27: Stakeholder views about stated preference methods



In the breakout group discussions, several participants mentioned that it was difficult to understand the three options for the stated preference format and it would be helpful to be guided by experts.

This is especially the case of Option 3 (impact-based approach). Many participants were not familiar with this approach. It was not clear to some, for example, how to translate the impact levels into relative impacts (across the sample) and how to translate relative impact into relative monetary values. It was also not clear to some participants how the approach would capture the impacts that customers might care about but which fall on other people (i.e. altruism).

It was also mentioned that customers dislike all the negative impacts, and may be equally averse to them. If so, it is not clear if the differences in impacts shown in the survey can be meaningful for customers.

In reality, the impacts are often "bundled" according to same common cause (e.g. some event that causes discolouration, low pressure, and supply interruptions). It was not clear for workshop participants if each impact shows in the survey controls for the others.

One workshop participant thought that choice experiments are 'safer': it is known what the limitations of this method are - the method is tried and tested. Using the new impact-based approach, it is not clear, for example, what is the possible effect of not having the money associated with the impacts shown on the same screen as the impacts themselves.

There was also some concern that using a new method, comparing with what was used in PR19, may result into large movements in the estimated values, for some companies.

Participants also suggested that, for the method to be successful, the description of the impacts needs to be clear and meaningful for customers (but still related to companies performance).

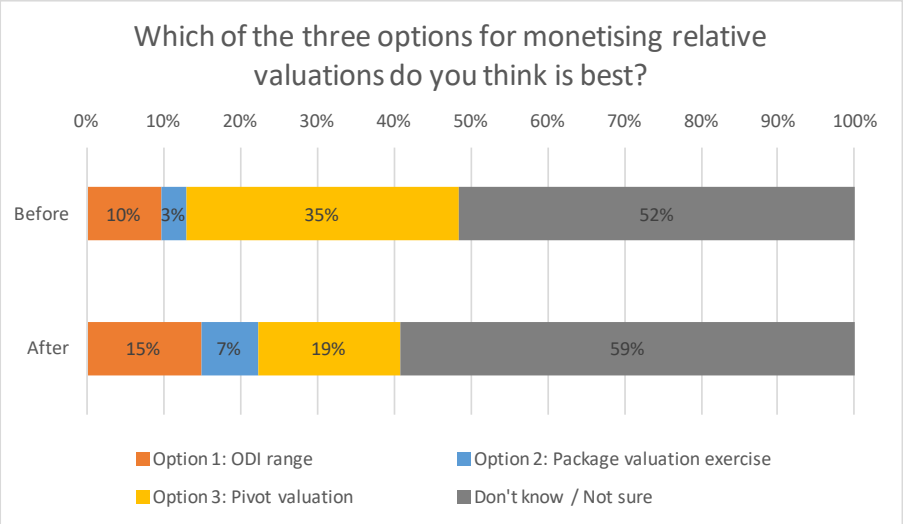
Despite these challenges, there was consensus that impact-based approach is potentially more robust that the other two options.

There was little discussion in the breakout group discussions about the other two options. However, it was mentioned that, in the menu/slider approach, it was difficult to derive consistent results on small impacts and respective costs. The method may this produce spurious accuracy.

Approach for monetising the value of impacts

The figure below shows the views of workshop participants about monetisation. Option 3 (pivot valuation) was the preferred option, out of the three options presented. However, the majority of participants stated they did not know which option was the best.

Figure 28: Stakeholder views about monetising



Most of the discussion in the workshop was about Option 3 (pivot valuation).

One participant mentioned that framing the Option 3 exercise around compensation feels a bit 'distant' from the ODI context of the research (unlike Option 1, which is directly related to ODI rates and can be used directly to inform policy choice).

Option 3 also assumes a minimum threshold of compensation to be meaningful for the customer. However, there are also other behavioural factors involved in the willingness to accept compensation, apart from the monetary value of the service that is lost: for example, the hassle of claiming compensation. In addition, customers may perceive immediately the utility of the compensation money but not of the impact. It is also not clear if the approach assumes that willingness to pay and willingness to accept are the same (i.e., if the context an improvement of deterioration of service)

These issues mean that the approach requires cognitive testing, to confirm that it derives the correct values.

Pivot valuation also relies on having a correct estimation of the relative impacts.

It was also agreed by several participants that the exercise should be anchored on more than one type of impact.

A final concern was about context. For example, the impact of 6 hours of supply interruption depends on the time of day (if those 6 hours are at night time, the impact is slight. The question should specify the impact and explain the context - however, this increases survey time.