Review of Impact Utilities (2018) 'Willingness to Pay Research to Support PR19'

A Note by Dr Paul Metcalfe, PJM economics, May 2018

1 Introduction

South Staffs Water and Cambridge Water (SSC) commissioned Impact Utilities to complete a stated preference (SP) study with the primary objective to estimate willingness to pay (WTP) for service level changes as a contribution to PR19 business planning. This note reviews the final report from the study.

2 Survey Design and Development

The report correctly identifies that an innovative approach was to be desired rather than a simple re-hashing of the PR14 design due to issues that had been raised in relation to that approach. The key challenges raised following PR14 included the need to ensure that customers fully understood the investment options that were presented to them, the need to ensure that all customer segments were engaged, and the need to triangulate results with other data sources to build confidence in the valuations. The report documents what appears to have been an appropriately thorough process of survey development to address these challenges involving several stages: stakeholder consultation, peer review of the methodology by Dr Ariel Bergman, deliberative research with household and non-household customers and a quantitative pilot phase.

The deliberative research phase appears to have been thoroughly conducted, including six focus groups and 10 depth interviews across a wide range of customer profiles. Details of the topic guide and findings from this stage are omitted from the main report, but the summary of findings suggests that they generated some sensible recommendations for survey refinement.

The pilot phase was also admirably used to quantitatively test a range of variations to the survey instrument to explore alternative approaches to that used for the PR14 survey (Eftec-ICS, 2013) and thereby arrive at a better presentation and design from a customer perspective. In total, four alternative approaches were tested, with the 'Future Outcomes' approach indicated as being preferred.

Following all this development work, the general valuation methodology remained similar to the PR14 approach insofar as discrete choice experiments involving choices between alternative service packages were used as the key elicitation method. The key innovations introduced related to the extent to which the wording/context of service levels was varied across the sample rather than using the same wording/context throughout, and the introduction of a new 'MaxDiff' exercise to obtain additional valuation evidence. An additional change from PR14 was to drop consideration of service deteriorations from the survey. This meant that only three levels were now considered for each measure: current (S0), some improvement (S1) and significant improvement (S2).

With regard to the wording/context of service levels, the differences related to whether they were presented with a 'public' or a 'private' focus. The public focus was, in general, in line with the PR14 form of presentation and took the form 'X out of X households per year'. The private version instead took the form ''Once in every X years' with respect to the personal chance of experiencing the issue. Mathematically, the service levels were the same in each case and it is an empirical question whether the difference in wording makes any difference to the values. Typically, WTP studies have not explored sensitivity to variations such as this and so the results from this study are of interest with regard to the methodology from a wider perspective as well as to SSC and its stakeholders in gaining a deeper understanding customers' preferences for PR19.

The 'MaxDiff' exercise was a second innovation introduced in this study in comparison to PR14. In contrast to a discrete choice experiment where customers choose between service packages, here they chose which service improvement they would like to see given the highest priority and which they would least like to see given the lowest priority from a set of five in each of a sequence of questions experimentally designed to cover the same full range of service measures and improvement levels included in the main discrete choice experiment.

The presentation of the MaxDiff exercise in the report is downplayed, treated as a supplementary piece of supporting evidence rather than as a comparable means of obtaining relative priorities between different types of improvement. Although the MaxDiff methodology is often an effective and valid technique of preference measurement, in the present case the key downside is that it does not explicitly show the bill impact of the options and so it is not clear whether participants were considering the options independent of their potential impacts on the bill, or if they were imagining how much each option would cost and mentally factoring that impact into their decision. I would imagine the former interpretation to be more likely, in which case the results obtained would represent relative WTP; however, the uncertainty over interpretation means that the measure has an additional source of error within it that the choice experiment results do not have. Overall, I would therefore support the interpretation of the MaxDiff results as being secondary in importance to the results from the choice experiment.

In most other respects, the design of the survey has generally been expertly carried out. The large set of service measures has been sensibly split into three groups: water quality, reliability of supply and environment, with separate choice experiments designed for each group. This is important because survey participants cannot typically trade off more than six or so attributes at a time so it is not wise to include all of them into one exercise. Moreover, it is good practice to create groups of similar attributes when splitting up the full set to make it easier for respondents to trade them off against one another.

The experimental designs for each choice exercise were created as 9 blocks of 13 scenarios of which 12 were random and 1 was fixed for everyone. In general, best practice would have been to use an 'efficient' design algorithm rather than a random design. (See Hoyos, 2010, for example.) The downside of a less efficient design is weaker statistical precision but there is no validity concern with the random design approach given sufficient number of blocks, which there were in this case.

Each participant completed only one of the three choice exercises and, unlike at PR14 and as recommended by UKWIR (2010) and UKWIR (2011), there was no package or contingent valuation exercise to value packages containing all service measures from all three groups. By contrast, the Eftec-ICS (2014) study for PR14 included a similar number of service measures, again split into three exercise, but asked each participant to complete all three exercises plus a package exercise combining all three blocks of service measures together. This approach was typical for PR14 WTP studies and also for PR19.

The omission of a package choice exercise is, in my view, a significant weakness of the study and stands it apart from most other WTP studies across the industry. The reason for including a package exercise is because the value of a package of service improvements tends to be less than the sum of its individually valued parts. In the absence of package values with which to scale (or constrain) the values obtained from each of the three individual discrete choice experiments it is possible that the package supported by cost-benefit analysis using these numbers will entail a bill impact greater than customers would be willing to pay as a whole for that package.

Whether or not this is an issue in the present case depends on the size of the package that is ultimately proposed in SSC's business plan. If this package, for example, is achievable without raising bills then there is no issue at stake. However, if a substantial bill increase is proposed, based in part on the WTP estimates obtained in this study, then there should be serious concerns about the validity of this increase. If this is the case then I would recommend SSC either undertake an additional piece of research to explore customers' maximum WTP for a broad ranging package of improvements (as recommended in the 'Next steps' section of the report) or, if this is not possible, apply a conservative approach by dividing all WTP values by 2, based on the fact that package-scaled values are typically approximately half the size of unscaled values from choice experiments. (This is loosely based on my having reviewed the majority of PR14 and PR19 WTP studies across the industry for the purposes of producing the Accent (2014) and Accent-PJM (2018) industry WTP comparison studies.)

3 Survey Administration

The survey was administered on a good-sized sample of households and non-households in the SSW and Cambridge regions. Household data was weighted to the local populations using Census 2011 data by age, gender and SEG.

The study also admirably focused on ensuring that hard-to-reach groups were adequately captured within the sample. One can therefore be confident that sampling error is sufficiently minimised for households.

With regard to non-households, it is not clear whether or not the sample is representative of the population as no population statistics are reported. Nor is it clear whether weighting has been applied to correct for any differences between the sample and the population. It would be helpful to include this information in the report.

The survey methodology comprised a mixture of online and face-to-face interviews. Online interviews are more cost-effective but there are limited numbers available for interview via panels, and they also do not capture some hard-to-reach groups. Supplementing online with face-to-face is therefore a sensible strategy to ensure full coverage of the population and sufficient sample sizes for all quota.

For non-households, there is the risk that online interviews are not completed by the appropriate person within the organisation. For example, a busy director may ask his PA to complete the survey rather than complete it himself. It would consequently be helpful if the report showed the proportion of non-household interviews that were completed online versus face-to-face. If any steps were taken to ensure that the correct person in the organisation completed the interview then these should be described. Otherwise, it should be acknowledged that online interviews may not completed by the appropriate person within the organisation.

4 Analysis Methods

The principle method of analysis of the choice experiment data is the Hierarchical Bayes (HB) logit model. This is consistent with best practice and the econometric modelling, in general, seems to have been expertly conducted.

5 Validity Appraisal / Data Triangulation

Typically, WTP studies are assessed according to two sets of validity criteria: content validity and construct validity. (See Bateman et al., 2002, for details.) Content validity is assessed based on expert review of the study design to ensure it is fit for purpose, and feedback data from the survey with regard to participants' understanding of the exercises, their ability to make meaningful choices, and the extent to which the scenarios presented were perceived as realistic. Construct validity, by contrast, is assessed by examining the consistency of results with expectations. This includes within-survey correlation analysis, and comparisons of the results against other studies.

With regard to content validity, the study can be considered to generally perform well in most respects, on the basis of points made elsewhere in this review. However, it is a notable omission from the study report that there are no feedback measures from participants with regard to understanding, ability to make meaningful choices, etc. Such questions are recommended in UKWIR (2011) for inclusion in WTP surveys for the purpose of appraising content validity.

With regard to construct validity, the first check is to confirm the basic expectation that people prefer more service improvement to less and prefer lower bills to higher bills. These expectations are satisfied by the choice experiment models. However, in the case of the MaxDiff results, the report states that there was little variation in priority whether the 'some improvement' or 'significant improvement' level was shown. This finding suggests, as the report rightly states, that participants were likely thinking in general terms about the priority of the service measure and did not evaluate the significance of the improvement range offered. This finding was not observed in the choice experiment models, and therefore represents an additional reason to attach more weight to the choice experiment results than to the MaxDiff results.

Further evidence with respect to construct validity is given in helpful sections in the report on the consistency of the MaxDiff and choice experiment results, and on data triangulation against findings from other companies. The results show that the MaxDiff and choice experiment results were indeed consistent with one another in general, which is supportive of the validity of the valuation results.

With regard to triangulation against external evidence, the approach taken is to compare values for 'some improvement' across studies without, it seems, any adjustment for differences in the amount of improvement offered. The underlying assumption appears to be that the scope of improvement offered is not relevant to the comparison. However, this scope of improvement is fundamental to the calculation of a unit value which is the value used for appraising what extent of any improvement should be undertaken, and what the ODI rate per unit should be. Because of this, the triangulation against external evidence in this report appears to me to be unreliable.

6 Grey Water Services

In addition to the main SP exercises, the survey also included a dichotomous choice contingent valuation question relating to willingness to pay for a grey water system. The reported WTP values of £5.12 / £4.12 per household per year for SSW and Cambridge customers respectively seem consistent with the frequencies of responses by bid level shown. However, it is not clear in the report how they have been derived. With dichotomous choice data an assumption is needed regarding the type of distribution of responses (e.g. normal, lognormal) or alternatively the Turnbull non-parametric method may be used, which is assured to give a lower bound on the mean. It would be helpful for the report to include details of the method used and to report the error range around the mean. This is particularly important if investment in this area is to form part of the company's PR19 business plan.

7 Conclusions

Overall, the study has generally been expertly conducted and best practice has been followed in most respects. Additionally, the study has innovatively included, and tested, different versions of the survey with respect to the presentation of service levels, including both a private and public form of presentation. The results from this comparison are interesting.

However, there is one key area of concern which is the absence of any package scaling factor to account for the fact that each set of service improvements will be delivered in the context of a broad ranging package of service measures at PR19. The study has not followed UKWIR (2010) and UKWIR (2011) guidance in this area and has not offered a justification for departing from this guidance.

Whether or not this is an issue in the present case depends on the size of the package that is ultimately proposed in SSC's business plan. If this package, for example, is achievable without raising bills then there is no issue at stake. However, if a substantial bill increase is proposed, based in part on the WTP estimates obtained in this study, then there should be serious concerns about the validity of this increase. If this is the case then I would recommend SSC either undertake an additional piece of

research to explore customers' maximum WTP for a broad ranging package of improvements or, if this is not possible, apply a conservative approach by dividing all WTP values by 2, based on the fact that package-scaled values are typically approximately half the size of unscaled values from choice experiments. (This is loosely based on my having reviewed the majority of PR14 and PR19 WTP studies across the industry for the purposes of producing the Accent (2014) and Accent-PJM (2018) industry WTP comparison studies.)

Notwithstanding this issue, SSC can have confidence in the results as presented in the main report as estimates of customers' WTP for service improvements for PR19.

8 References

Accent (2014) Comparative Review of Willingness to Pay Results, Final Report, June 2014.

Accent-PJM (2018) Comparative Review of PR19 WTP Results, Draft Report, May 2018.

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

Eftec-ICS (2013) South Staffs Water PR14 Stated Preference Study: Final Report, June 2013.

Hoyos, D. (2010) The state of the art of environmental valuation with discrete choice experiments, *Ecological Economics*, **69**, 1595-1603.

Turnbull, B. (1976) The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data, *Journal of the Royal Statistical Society. Series B (Methodological)*, **38**(3), 290-295.

UKWIR (2010) Review of Cost-Benefit Analysis and Benefit Valuation, Final Report and Practitioners Guide (Part 1 – Benefits Valuation), Report Ref. No. 10/RG/07/18.

UKWIR (2011) Carrying Out Willingness to Pay Surveys, Final Report and Practitioners Guide, Report Ref. No. 11/RG/07/22.