Review of Impact Utilities (2018) 'Willingness to Pay Research to Support PR19 – Retest + Focus on Retail Attributes'

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

1 Introduction

South Staffs Water and Cambridge Water (SSC) commissioned Impact Utilities to complete a second stated preference (SP) study (Wave 2), following on from the first (Wave 1) which was reviewed recently[1]. The primary objective of the Wave 1 study was to estimate willingness to pay (WTP) for service level changes as a contribution to PR19 business planning. The Wave 2 study differed from the first in a few key respects – as discussed in this note – but had the same aim of estimating WTP for service improvements.

This note reviews the final report from the Wave 2 WTP study.

2 Survey Design and Development

The Wave 2 survey was designed to test the sensitivity of WTP values to differences in attribute wording and service levels, to add retail service measures to the set of attributes tested, to test sensitivity to package effects, and to test whether a lower bill starting point with an improved level of service altered the WTP values.

The development of the survey sensibly included a deliberative research phase to test the changes in survey wording applied since Wave 1 and a good-sized quantitative pilot of 142 interviews split between household and non-household. This adheres to good practice.

2.1 Key Differences Between Wave 1 and Wave 2 Surveys

In many respects the Wave 2 survey instrument was similar to the Wave 1 survey. The key differences were the following.

- The up-front Grey water recycling questions were removed.
- The MaxDiff exercise was removed.
- Each participant saw two choice experiment exercises rather than just the one, plus a package choice question. The second choice experiment contained a new set of attributes focused on retail attributes, while the first focused on one of the three groups of attributes also included in the Wave 1 survey.
- Within the three groups of attributes carried forward from Wave 1 to Wave 2, three individual attributes were excluded from Wave 2: Drought restrictions, Giving customers control of their water usage and Traffic disruption.
- For the attributes that were carried forward from Wave 1 to Wave 2, all attributes were worded with a public orientation rather than a split sample between of public and private a number of changes were made to the wording.

[1] Filename: 3204_WTPreview_PJM.docx

    - **Water not safe to drink**: duration of incident increased from 2 to 3 weeks.
    - **Lead pipes:** altered from no health risk to almost none.
    - **Water hardness**: definition altered to state that hardness is good for health.
    - **Unexpected temporary loss of water supply**: duration of incident changed from 'up to 24 hours' to '1-5 hours'.
    - **Managing impacts on rivers and streams**: a more detailed and descriptive description shown in Wave 2.
- In addition to these wording changes, there were changes to the service levels shown for most of the attributes. In most cases, the changes involved reducing the scope of improvement shown quite substantially. In particular, for the following attributes the 'S2' substantial improvement in Wave 1 was set equal to zero whereas the Wave 2 S2 level was greater than zero:
    - Water not safe to drink
    - Discoloured water
    - Taste and smell of water
    - Lead pipes
    - Unexpected temporary loss of water supply
    - Temporary use ban
    - Low water pressure
    - Flooding from a burst pipe
- Finally, a split-design was created, for households only, whereby one group saw a version with a lower bill presented as the starting point (the 'Low bill' scenario). All other households, and all non-households were shown a scenario where the starting point was equal to the current bill level (the 'Current bill' scenario).

These changes can be expected to have influenced the values obtained from the survey instrument in a number of ways. One important effect is that, in general, we would expect reductions in the scope of service change offered to lower the value for the change in service level offered, in comparison to the previous wave, but by a less-than-proportionate amount. Thus if the scope of service change offered halved for an attribute we would expect the WTP value to fall but by less than one half. (PJM-Accent, 2016).

However, in cases where the initial maximum (S2) service improvement entailed a reduction to zero in the number of service issues, the WTP value may possibly fall by more than a proportional amount due to the special significance of 'zero risk' as a driver of choice (Kahneman and Tversky, 1979; Schneider et al. 2017). In general, people tend to give excessively high value to removal of the last trace of risk to go from a very low risk to a certainty. If the value for a range of service improvement with an end-point of zero is then linearly interpolated over the range to derive unit values then these unit values will tend to give too high a value, in general, except for the very last increment to get to zero in which case the interpolated values will understate the true feelings of respondents.

The implication of this is that unit values should be higher in Wave 2 than in Wave 1 where there are reductions in the scope of service change offered and where the Wave 1 service level in question was not equal to zero. This is because unit values are

derived by dividing the value for the service level change through by the amount of service change offered. Thus if the value for the service level change decreases by less than the amount of service change offered, the unit value will increase. In cases where the Wave 1 service level in question was equal to zero, and the corresponding Wave 2 service level was greater than zero, the unit values themselves may reasonably be lower in Wave 2 than the unit values from Wave 1 due to the special significance of zero.

Offsetting, or enhancing, these effects are the expected impacts of the various changes to the attribute wording. For example, we would expect values for *Water not safe to drink* to rise due to the increase in the duration of the incident from 2 weeks to 3 weeks; we would expect values for *Lead pipes* and *Water hardness* to fall because of the changed health impacts; we would expect the values for *Unexpected temporary loss of water supply* to fall due to the change in the duration of interruption shown; and we would expect values for *Managing impacts on rivers and streams* to rise due to the enhanced descriptions shown in this case.

With regard to differences between the Low bill and Current bill scenario, in general we would expect values to be higher under the Low bill scenario due to the fact that people tend to be more sensitive to bill increases than bill reductions.

These differences, and their implications, are important for assessing the stability of underlying values from Wave 1 to Wave 2, and thereby for assessing their reliability. They are also important for assessing which values are appropriate to take forward, via triangulation, to CBA and ODI rate-setting.

3 Survey Administration

The main stage of the survey included online and face-to-face interviews, as did the Wave 1 survey, and the full sample, combining main stage and pilot, contained 982 interviews in total split across households and non-households in the South Staffs (SSW) and Cambridge (CAM) areas.

In general, these sample sizes are adequate; however, the sample was skewed towards the SSW area and towards households, meaning that only a small number of non-household interviews were completed in the CAM area. The split between SSW and CAM areas for the non-household sample is not given in the report; however, the accompanying Excel workbook indicated that there were only 57 non-households interviewed in the CAM area. In itself, one might expect this to return reasonably precise results. However, each respondent saw only one of the three original attribute blocks (Quality, Security, Environment) and so there were, on average, less than 20 participants who completed each block of choice questions in this segment. On this basis, there is an insufficient sample size, in my view, to have confidence in the non-household WTP results for the CAM region.

By contrast, the household sample for SSW and CAM regions and the non-household sample for the SSW region seem to be of adequate size to deliver results. With regard to the household sample, the original structure matched well to Census statistics by age, gender and SEG, and the final results are in any case weighted to these Census statistics and so thereby reasonably assured to be representative. It is less clear

whether the SSW non-household sample is representative given that no statistics are reported for the population against which to compare the sample. The report states that Impact 'took steps to ensure that we spoke to a sample that broadly reflected the total population: we aimed to achieve a mix of business sectors and company sizes that were broadly in-line with a random sample of NHH leads provided by SSC.' The degree of representativeness is not thereby assured to the same extent as for households. Given that sample sizes are also smaller for non-households than for households, one should therefore be less confident in the non-household results, in general, than the household results.

4 Survey Feedback

Feedback to the survey from participants was positive, and this should give confidence in the ability of the survey to generate meaningful results.

5 Analysis

The method of analysis appears to be consistent with the Wave 1 survey via use of the Hierarchical Bayes (HB) logit model. This is consistent with best practice.

6 Findings

6.1 Comparison of Wave 2 and Wave 1 Results

The report provides some comparisons between Wave 1 and Wave 2 results, in terms of WTP per customer for improvements to S1 and S2, and in terms of standardized unit values. With regard to the WTP per customer values for S1 and S2 improvements, the reported comparison suggests some significant reductions in values, even when comparing Wave 2 S2 values against Wave 1 S1 values. These results are somewhat surprising, given the expectations set out in Section 2.1 of this note.

However, the Wave 1 values shown are based on the combined Public and Private values. In my view, the appropriate set of Wave 1 values to compare against is the 'Public' set of values only. This is because there were found to be differences in Wave 1 between Public and Private valuations and the Wave 2 survey replicated the Public formulation. The public values tended to be lower than the Private values and so part of the difference observed between Wave 1 and Wave 2 is accounted for by this difference.

It would be helpful to show this comparison in the report as part of an enhanced examination of the causes of the differences between Wave 1 and Wave 2, to the extent that these can be determined. In general, an inability to adequately explain the sources of the differences between survey waves will undermine the perceived reliability of the results.

With regard to unit values, the results are difficult to interpret in their current form due to the fact that non-household values are combined with household values and SSW values are combined with CAM values. I would recommend that the values are presented alongside one another for each segment (HH/NHH; SSW/CAM), and with the Public Wave 1 values being used for the calculation. This will allow for a more

meaningful comparison and understanding of the differences and their causes in light of the fact that non-household values are less reliable than household values, particularly in the CAM area where the sample size was very small.

On the basis of the numbers presented, it seems that there are some very substantial differences in some cases, both increases and decreases. These differences are not sufficiently well explained in this report to give confidence in the results as being reliable. This should be possible to rectify to a certain extent via an enhanced examination of the differences found in light of the considerations set out in Section 2.1 of this note, alongside any further views that Impact has regarding the causes of the differences.

7 Sensitivity to Initial Bill Level

The report includes an analysis of the difference in values obtained between those seeing the Low bill scenario and those seeing the Current bill scenario. The results here run counter to expectation, both mine and the authors', in that the unit values are lower under the Low bill scenario rather than higher. The explanation given by the authors is plausible, however, and I support the recommendation to treat the Low bill scenario as a sensitivity test while using the Current bill scenario as the main case.

8 Package Effects

The Wave 2 study included a Package exercise, in line with best practice (UKWIR, 2010, 2011), to explore how sensitive values are to the size of the package being valued. The design of the exercise took the form of a single-bounded dichotomous choice contingent valuation question followed by an open-ended contingent valuation question both to value a package containing either 2 or 3 sets of attributes all improving from S0 to S2.

The sample was split between those that saw packages of 2 blocks of attribute improvements and those that saw 3 blocks. However, it is not made clear in the report how the blocks were allocated, and no results are given that depend on the blocks that are being valued. This is significant because the valuation results suggest that there are substantially higher values for some blocks than others. In particular, there is a big difference between values for Retail attributes as a whole and those for the Security and reliability attributes.

The results suggest that, as expected, 3-part packages are more valuable than 2-part packages. (However, we don't know what these packages include.) The results also show that only around 40% of households and non-households chose the 2-block improvement package at the lowest cost level offered. Ideally the design would have included a £5/5% level to identify the lower tail as, given the data at hand, it was necessary to extrapolate beyond the data observed to estimate the likely point at which 50% would prefer the improvement package, ie the median acceptability. No estimate is given for the mean value, which is the usual value taken forward for CBA, and which tends to be higher than the median.

Nonetheless, the results in Table 8 of the report are in line with expectation for households. For non-households the values suggest scaling factors below expectation,

and below what could be considered reasonable. This is because the value of the 2-block package was found to be less than half the DCE-derived value. This indicates an inverse sensitivity to scope, ie the more the improvement the lower the value, which is clearly not valid. One possible reason for this is that the median value is used for the package exercise rather than the mean. I would suggest that the Turnbull method is used to calculate a lower bound estimate of the mean package value, which may overcome this issue with the non-household results.

While the package exercise is an important addition to the survey since Wave 1, it is not necessarily the case that the DCE results should be scaled to be consistent with a package value. To do so could be argued to be excessively conservative if it is the case, as we understand it to be, that the package that is ultimately to be included in SSC's business plan is likely to entail falling bills in real terms rather than rising bills. However, if a substantial bill increase is ultimately proposed, based in part on the WTP estimates obtained in this study, then the WTP results should ideally be scaled to be consistent with the package exercise results.

9 Conclusions

Overall, the Wave 2 research has been conducted to a high standard, in line with the Wave 1 research. The study has usefully tested the sensitivity to package size effects and has incorporated important revisions to the wording of the service measures and the service levels adopted.

There are two key areas that I would draw attention to: firstly, the non-household sample size for the Cambridge area appears to be too small to draw reliable inferences from; and secondly, the comparison of Wave 1 and Wave 2 WTP results warrants some further examination and explanation in the report. Otherwise, SSC should have confidence in the results as presented in the main report as estimates of customers' WTP for service improvements for PR19.

10 References

PJM-Accent (2016) Analysis of PR14/SRC15 WTP Findings, A Report for Scottish Water, 24 September 2016.

PJM (2018) Review of Impact Utilities (2018) 'Willingness to Pay Research to Support PR19', A Note by Dr Paul Metcalfe, PJM economics, May 2018.

Kahneman, D. and Tversky, A. (1979) Prospect Theory: An analysis of decision under risk. *Econometrica*, **47**, 263–292.

Schneider, E. Streicher, B., Lermer, E., Sachs, R., Frey, D. (2017) Measuring the Zero-Risk Bias, *Zeitschrift für Psychologie*, **225**(1), 31-44.

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.