Representations:

Assessing the Statistical Validity of C-MeX

Document Reference: J002b

This report presents the findings of Frontier Economics from their independent review of the statistical strength, and the pragmatic options for improving, the proposed methodology underpinning the new C-MeX incentive.

United Utilities Water Limited

United Utilities
helping life flow smoothly

An introduction by United Utilities

We welcome the clarity that Ofwat's publication of revised C-MeX guidance as part of the draft determinations outcomes appendix provides. We continue to endorse the objectives and overarching structure of the C-MeX proposals, and recognise the crucial role this new incentive will have on focussing industry efforts on boosting service for customers over the coming years.

Ofwat has described C-MeX as a key component of the AMP7 regulatory regime, and it is crucial that it commands the confidence of customers, companies and wider stakeholders. However, we have some concerns with the latest proposed design and operation of the C-MeX survey. We therefore requested that Frontier Economics review the statistical validity of the C-MeX mechanism, as set out by the latest proposed methodology. We also asked that they independently consider pragmatic options to address any concerns they may find. The results of their work is presented in the following report.

Frontier have shown that the current detailed design of C-MeX introduces a high degree of statistical uncertainty in results. They have also identified a range of relatively small changes to methodology, survey structure and incentive calculation that can materially address some of the limitations with the existing proposed methodology. In particular they have highlighted the following opportunities:

- **Double C-MeX sample sizes:** Ofwat should at least double sample sizes to get the confidence intervals that are referenced in the C-MeX methodology paper. The approach to defining separate sub-pots for telephone and digital contacts is further reducing statistical confidence levels and should be reviewed.
- **Discontinue the use of Net Promoter Score (NPS):** The use of NPS is actively reducing statistical significance of results, and is not in practice adding stretch to C-MeX as it is highly correlated to satisfaction scores. There are clear indications that customers do not understand how NPS ratings differ from general satisfaction ratings. Removing NPS from C-MeX calculations will help boost the statistical significance of final company scores substantially.
- **Consider how the sample size is divided between the customer service and experience components:** The customer service survey has a larger impact on the uncertainty of results than the customer experience survey, as customer responses are spread across a wider range. Allocating a larger percentage of available survey size to the customer service survey will therefore help reduce the overall level of noise.
- **Remove 'cliff edges' from incentive calculations**: The way in which Ofwat currently propose to apply financial and reputational incentives is not supported by the uncertainty ranges observed for annual company scores. A more gradual financial incentive, with fewer cliff edges, would be preferable. Similarly, ranking individual companies 1-17 in end of year performance reports is not supported by the data. Instead, grouping companies into performance bands (good performers, average performers, poor performers, etc.) would be more appropriate given the uncertainty of data.

The issues Frontier have identified, if left unaddressed, risk undermining confidence in C-MeX scores and incentives. However, all of these issues can be directly and efficiently addressed, without requiring major changes to the established C-MeX methodology.

ASSESSING THE STATISTICAL VALIDITY OF C-MEX

A report for United Utilities

30 August 2019

Annabelle Ong
020 7031 7056
annabelle.ong@frontier-economics.com

Katharine Lauderdale
020 7031 7000
katharine.lauderdale@frontier-economics.com

CONTENTS

EXECUTIVE SUMMARY

At PR19, Ofwat is replacing the previous customer satisfaction measure (SIM) with a new “Customer Measure of Experience” (C-MeX), to be effective from 1st April 2020. Ofwat’s Draft Determination has signalled that C-MeX will be a composite measure that is based on:

- customer satisfaction (customers that have been in contact with the company)
- customers’ experience (general customer base); and
- Net Promoter Score (included in the satisfaction and experience surveys)

C-MeX scores are associated with significant financial incentives, with rewards up to 6% (12% in some circumstances) and penalties up to 12% of residential retail revenues. Additionally, companies have a reputational incentive to perform well with regard to C-MeX, as the annual scores and rankings are published and customers and stakeholders care about companies’ performance.

United Utilities has asked Frontier Economics to assess the statistical validity of the current C-MeX design. Statistical validity is important: companies need precision in the rankings and resulting penalty and reward payments in order for the financial and reputational incentives to be fully effective, and also in order for C-MeX to be a useful diagnostic tool.

This report assesses the following questions:

- To what extent can we be confident that the differences between C-MeX scores reflect actual differences in companies’ performance?
- What sample sizes would be required to be confident in differences in companies’ scores?
- What are the practical steps that Ofwat can take to improve statistical validity and confidence in C-MeX?

We have conducted an independent and objective assessment based on statistical methods, and find the following:

- Companies cannot be confident that their performance is different from companies that have received different reward or penalties from them. Given Ofwat’s current sample size, we cannot be confident that there is a real performance difference between companies with annual scores less than 2.7 points apart.
- C-MeX has limited precision as a diagnostic tool for a company’s own performance. A company that has seen a year-over-year change in their overall score of less than +/- 2.7 points, or a quarter-over-quarter change in their overall score of less than +/- 5.4 points, cannot be confident that this is due to a real performance change.
- Companies cannot be confident in their ranking. We estimate that a company faces a chance of having received the incorrect rank that ranges between 40% and 75% under the proposed design. Companies in the middle of the distribution have particularly high risk of having been mis-ranked.

- Companies face a high risk of having been allocated to the incorrect reward or penalty bucket. For example, a company assigned to the Penalty bucket under the proposed design has a 37% chance that it should have received a different allocation:

A company assigned to the payment bucket:	Has the risk of having been allocated to the incorrect payment bucket:
Highest payment	17%
Payment	35%
No payment	36%
Penalty	37%
Highest penalty	11%

- A company that fails to achieve its target sample size will increase the risk of mis-ranking and misallocation of incentive rewards and penalties for all companies. This is because every company's score contributes to every other company's ranking and calculation of reward and penalties.
- Companies will not receive the correct average payment or penalty in the long run. The financial cost of being incorrectly misallocated to a lower performance category will not on average cancel out the financial benefit of being incorrectly misallocated to a higher performance category. This is because the rewards and penalties are asymmetric and have cliff-edges.
- Under the proposed design, companies will likely experience substantial random fluctuations in their ranking and payments. This may erode confidence in the scheme, and lessen the reputational and financial incentives over time.

We have investigated how the reliability of the rankings, rewards and penalties would improve with an increased sample size. We find that:

- A larger sample size would substantially reduce the risk that a company has been mis-ranked. The company in the proposed design with the highest mis-ranking risk, at 75%, would see this risk reduced to 67% with a doubled sample size, or 60% with a quadrupled sample size. Even with a quadrupled sample size, the risk of mis-ranking is still very high.
- A larger sample size would also reduce the risk that a company has been allocated to the incorrect reward or penalty bucket. The bucket with the highest risk of misallocation, at 37%, would see this risk reduced to 31% with a doubled sample size, or 24% with a quadrupled sample size. Even with a quadrupled sample size, the risk of misallocation is still substantial.

Using data from United Utilities, we have also investigated other practical steps that Ofwat could take to increase the reliability of C-MeX:

- The NPS measure contributes heavily to the noise in the overall score. This is for two reasons. First, the way the NPS score is calculated from its survey question introduces noise. Second, NPS is highly correlated with the other C-MeX components. We estimate that dropping NPS from the overall score would reduce the standard error (level of noise) by 27%, which is the improvement one would expect from increasing the sample size by nearly 90%.

- Using pilot and shadow year data, Ofwat could refine the allocation of sample size between different components. We find that a higher precision in the overall score could be achieved by redistributing survey resources, holding the sample size fixed. We estimate that both omitting NPS and redistributing survey resources from customer experience to customer satisfaction would reduce the standard error by around 30%, which is the improvement one would expect from increasing the sample size by nearly 100%. Ofwat may be able to achieve a higher increase in precision by further considering the allocation of resources within the different sub-components of the surveys.
- We estimate the combined effect of our recommendations: removing NPS, reallocating sample sizes between components, and doubling the sample size. Even with these changes, there would still be a material risk of mis-ranking, ranging from 22% to 65% for individual companies. There would also still be a risk of having received the incorrect reward or penalty: we estimated that the ‘payment’ bucket would have the highest associated risk of misallocation under the proposed design, at 29%. Therefore we also suggest that Ofwat consider changes to the design of the reward and penalty buckets.
- The cliff-edges create large financial risks to misallocation. By removing the cliff-edges from the incentive design and moving to a more gradual scale, the total monetary amount that may be misallocated would be reduced.

To make a final decision on C-MeX, we recommend that Ofwat replicate our analysis using data from all companies. This would allow Ofwat to make informed choices about the final sample size, the allocation of sample size to components, and the design of the scheme.

1 BACKGROUND AND OBJECTIVE

Background

At PR19, Ofwat is replacing the previous customer satisfaction measure (SIM) with a new “Customer Measure of Experience“ (C-MeX). Whereas SIM was focused on the satisfaction of those customers who contacted each water company and the number of contacts, Ofwat’s Draft Determination has signalled that C-MeX will be a composite measure that is based on:

- customer satisfaction of those customers that have been in contact with the company (40% weighting);
- customers’ experience of the general customer base (40% weighting); and
- Net Promoter Score (NPS) from customers that have been in contact with the company and also the general customer base (20% weighting).[1]

C-MeX will be effective from 1st April 2020. In January 2019, Ofwat published data on the pilot results for the C-MeX surveys and for 2019/20 Ofwat is running the survey as a “shadow reporting” year.

C-MeX scores are associated with significant financial and reputational incentives: Ofwat has indicated that companies that perform above one standard deviation from the mean will receive an uplift of 6% of residential retail revenues whereas companies below one standard deviation from the mean will pay a 12% penalty. Figure 1 below shows the full incentive structure Ofwat is proposing for C-MeX. The scores are based on data for the pilot year.

Figure 1 Ofwat’s proposed incentive structure based on pilot year data

	Score	SD	Type	Rate
1	84.4	1.5	Highest payment	+6.0
2	83.4	1.2	Highest payment	+6.0
3	83.2	1.1	Highest payment	+6.0
4	83.0	1.1	Highest payment	+6.0
5	82.7	1.0	Payment	+2.9
6	81.1	0.5	Payment	+1.4
7	80.8	0.4	Payment	+1.2
8	79.8	0.1	None	0.0
9	79.7	0.0	None	0.0
10	79.4	-0.1	None	0.0
11	79.3	-0.1	None	0.0
12	78.3	-0.4	Penalty	-2.4
13	77.2	-0.7	Penalty	-4.5
14	76.2	-1.1	Highest penalty	-12.0
15	75.5	-1.3	Highest penalty	-12.0
16	75.1	-1.4	Highest penalty	-12.0
17	74.1	-1.7	Highest penalty	-12.0

Highest payment for companies above 1SD above the mean

Companies between 1 and 0.25 SD above the mean: *SD score x 3% = payment*

Companies between 0 and ±0.25 receive no payment/penalty

Companies between -0.25 and -1 below the mean: *SD score x 6% = penalty*

Highest penalty for companies 1SD below the mean

Source: Ofwat, 2019, PR19 Customer Measure of Experience (C-MeX): Policy decisions for the C-MeX shadow year 2019-2020

[1] The NPS is a customer experience metric calculated based on the response to the question, “if you could choose your water provider, how likely would you be to recommend [Water Company] to friends or family?” (scale of 0-10). The NPS question is included in the customer service and customer experience surveys. Ofwat, 2019, *PR19 Customer Measure of Experience (C-Mex): Policy decisions for the C-Mex shadow year 2019-2020*

In addition to the financial incentive, companies have a reputational incentive to perform well with regard to C-MeX as the annual scores and rankings are published and customers and stakeholders care about companies' performance.

Project objective

Ofwat is currently proposing to base C-MeX on a sample size of 1,600 customers per company per annum (800 for the satisfaction survey and 800 for the experience survey). As C-MeX scores have significant reputational and financial consequences, the industry needs to be confident that the differences in C-MeX scores are representative of actual differences in companies' performance.

United Utilities has therefore asked Frontier Economics to assess the statistical validity of the current C-MeX design. Statistical validity is important: companies need precision in the rankings and resulting penalty and reward payments, both in order to use the customer satisfaction information to target changes in behaviour, and also in order for the financial and reputational incentives to be fully effective. If scores and rankings are subject to a large amount of noise, companies have little control over their C-MeX scores in any given year, which undermines the incentives.

The objective is to assess the following questions:

- To what extent can we be confident that the differences between C-MeX scores reflect actual differences in companies' performance?
- What sample sizes would be required to be confident in differences in companies' scores?
- What are the practical steps that Ofwat can take to improve statistical validity and confidence in C-MeX?

We have conducted an independent and objective assessment based on statistical methods.

Report outline

This report is structured as follows:

- Section 2 provides our assessment of the proposed sample size and resulting level of confidence in the C-MeX scores;
- Section 3 provides our detailed analysis of the sources of noise in the overall score and how to improve the level of confidence in C-MeX ; and
- Section 4 presents our recommendations.

We find that on the basis of the current sample size and survey structure companies cannot be confident that their rankings reflect actual differences in company performance. As a result, there is a high risk of misallocating rewards and penalties. This risk is exacerbated by the cliff-edges that are part of the incentive design. Ofwat can take simple steps to improve the reliability of the scores by removing NPS and by increasing the sample size. The risk of misallocation would also be reduced by removing cliff-edges.

For those who would like more detail on our analysis or wish to replicate our analysis, we have included a Technical Annex (Annex A) that describes our calculations in detail.

2 DOES THE PROPOSED SAMPLE SIZE PRODUCE RELIABLE C-MEX SCORES?

The purpose of this section is to address the following questions:

- To what extent can we be confident that the differences between C-MeX scores reflect actual differences in companies' performance?
- What sample sizes would be required to be confident about differences in companies' scores?

We first need to establish the level of noise in the scores

The C-MeX surveys aim to gather information on customers' views about the level of service that their water company provides. As we cannot ask all customers for their views, the survey has to be based on a sample of customers. The larger the sample size the higher the precision of the results and the more confident we can be that the score reported for each company truly represents the view on the company's performance across all of its customers. The level of noise in the scores is summarised by the *standard error*, which is a statistical measure of the precision or variability of an estimate.

To assess the specific sample size required for a particular level of precision, we first consider the precision that is achieved under Ofwat's proposed sample size. We note that, in moving from the 5-year Service Incentive Mechanism (SIM) review to an annual C-MeX assessment, Ofwat has effectively reduced the sample size it uses to evaluate customer satisfaction. Moreover, not only the precision from the annual sample but also the precision from the quarterly sample is important. The quarterly C-MeX results are a useful management tool for focusing immediate efforts at improvement, but they can only be utilised in this way if companies can be confident that they are not noisy and that they reliably reflect recent changes in customer satisfaction.

For our assessment, we calculated an implied standard error based on Ofwat's reported confidence interval for individual scores. Ofwat has stated that the current sample size of 1,600 per annum per company leads to a confidence interval of +/-1.9%.[2] Ofwat does not explicitly state the level of confidence in its report but based on previous statements around SIM and standard practice, it is not unreasonable to infer that Ofwat uses a 95% confidence interval. A 95% confidence interval would imply a standard error of 0.97.[3]

---

2 This confidence interval is provided in Ofwat, 2019, *PR19 Customer Measure of Experience (C-Mex): Policy decisions for the C-Mex shadow year 2019-2020.*

In their pilot study report for Ofwat, Navigator/Allto have stated that this confidence interval is based on the assumption that the true average in the customer population is 50%. Given that the C-MeX survey questions are on numeric scales, and are not elsewhere represented as proportions, we are unsure of the calculation underlying this confidence interval. We assume this means a confidence interval of +/- 1.9 on the 100-point C-MeX scale. Please see Navigator/Allto, 2019, *C-MeX Pilot for PR19: Report prepared for Ofwat*.

3 A 95% normally-distributed confidence interval width is equal to 1.96*(standard error).

In this report we have used estimates of standard errors for particular analyses based on the different advantages of the standard error estimates. In our analysis of the precision of the overall C-MeX score, we use the standard error that Ofwat has provided, because it is based on data from all companies. In our later analysis of the individual components of the C-MeX score, we estimate the standard error of individual components of C-MeX based on data from United Utilities.

We then need to assess which scores are statistically different from one another

As the reputational and financial incentives around C-MeX are based on relative performance and rankings, the relevant test is to consider how confident we can be that any two particular scores are distinguishable.[4] But the confidence interval for a single score does not answer this question, because it only includes the uncertainty in the score for a single company. A test that considers whether the scores of two different companies are significantly different from one another takes into account that there is some degree of uncertainty around both scores. Using the standard error quoted above, we can construct a test for this,[5] and assess whether any particular two scores are, in fact, distinguishable.

If we cannot say with a reasonable degree of certainty that two scores are different, then the incentive system runs the risk of misallocating rewards and penalties because two firms may receive different rewards/penalties despite the fact that the data is insufficient to be sure that their performance is not the same. In addition, companies will have a much reduced incentive to improve their scores as changes in the rankings may not necessarily reflect changes in underlying performance but instead be based on random variation in the survey results.

We cannot be confident that scores are distinguishable based on the current sample size

Figure 2 below shows the minimum gap between scores that we can be 95% confident about for various sample sizes, i.e. the gap in the survey scores that we need to observe to consider there is only a 5% chance that, in fact, the two companies have the same underlying performance. The figure shows that:

- Using Ofwat's current sample size of 1,600, the minimum gap between scores that we can be confident about (at the 95% certainty level) is 2.7. This means that we cannot be confident about any gap between scores that is smaller than 2.7.
- By way of example, in Figure 1 we see that Ofwat's pilot year data scores show that the scores for the 4th and the 5th company are 83.0 and 82.7. This gap is clearly smaller than 2.7 and therefore we cannot say with confidence that the difference in scores is based on actual differences in companies' performance. This means that the 4th company which would receive a reward of 6% is not distinguishable from the 5th company which would only receive a reward of 2.9%. As Ofwat's incentive design includes cliff-edges (where reward and penalty rates are subject to a step change between two scores) instead of gradual changes in incentives, the risk of misallocating a significant amount of penalty or reward is exacerbated.
- The minimum gaps in Figure 2 also reflect whether a company that observes a change in its score over time can be confident that there was a real change in

[4] For a company that is trying to judge whether its performance has changed over time by comparing its results across quarters or years, this is also the relevant test.

[5] We use the Z-test to assess whether two scores are distinguishable. For more detail, please refer to Annex A.

performance. In other words, a company that has seen a year-over-year change in their overall score of less than +/- 2.7 points cannot be confident that this is due to a real performance change and not just noise in the data. A company that has seen a quarter-over-quarter change in their overall score of less than +/- 5.4 points cannot be confident that this is due to a real performance change. This may not be sufficient precision in order for a company to use its score as a diagnostic tool.

- The confidence interval quoted by Ofwat above implies that Ofwat consider 1.9 points as an acceptable margin for error in the scores. Figure 2 shows that a 1.9-point gap between scores (at a 95% confidence level) would require a sample size of 3,200, not 1,600. The confidence interval only takes the uncertainty in one company's score into account, but we need to take the uncertainty in both companies' scores into account in order to compare them. Even though the confidence interval around a single score is not directly comparable to the results in Figure 2, this would mean that Ofwat would need to at least double the sample size.[6]

- We note that in fact many scores estimated for the pilot year are quite close together, as shown in Figure 1. Therefore, to be confident that rankings, rewards and penalties are based on actual differences in underlying performance, the sample size needs to be increased significantly.

Figure 2 Gap between scores that we can be 95% confident about for different sample sizes

Sample size	Gap between scores that we can be confident about (minimum statistically detectible difference between scores)
400	5.4
800	3.8
1,600 (Ofwat's current sample size)	2.7
2,400	2.2
3,200	1.9
4,000	1.7
6,400	1.3

Source: Frontier Economics; analysis of figures in 'PR19 Customer Measure of Experience (C-MeX): Policy decisions for the C-MeX shadow year 2019-2020' (confidence interval estimate on p12)

The currently sample size leads to a high risk of misallocating rewards and penalties

To illustrate how the current sample size can lead to misallocated rankings, we have used the scores shown in Figure 1[7] that are allocated into Ofwat's different

[6] We note that the confidence interval of +/- 1.9% is not directly comparable to the minimum statistically detectible difference. The confidence interval is based on a single score, while the 2-sample z-test is based on two scores.

[7] Although Ofwat stated that the scores in Figure 1 were based on pilot year data, we were unable to reconcile the scores with Navigator/Allto pilot year data (i.e. in Navigator/Allto, 2019, *C-MeX Pilot for PR19:*

penalty and reward buckets. Figure 3 shows, for each ranked company, the ranks of other companies whose scores are not significantly different at the 95% level.[8] For example, the company that is ranked 7th is not significantly different from the companies ranked 2nd to 12th. The rankings shaded in green show where a pair of companies with scores that are statistically indistinguishable fall within the same penalty and reward bucket. All of the red rankings illustrate where a pair of companies with scores that are statistically indistinguishable have fallen in different penalty and reward buckets.

Figure 4 shows the estimated probability that each company was assigned the incorrect rank.[9] For example, we estimate there is a 74% chance that the 7th rank company would have been assigned a different rank if we had been able to observe all companies' scores without any noise. The risk of mis-ranking each company ranges from 40% (company 17) to 75% (company 9). This shows that companies' scores are too close together to be able to reliably rank, given the level of uncertainty in the scores. Each company's rank depends not only on its own score, but also on the scores of every other company. As a result, the noise in one company's score will contribute to the risk of mis-ranking other companies as well, particularly other companies with similar scores.

We understand that currently the rankings create significant reputational incentives to improve customer satisfaction. However, there is a risk that the current level of noise will lessen that useful incentive, particularly over time. If companies experience random fluctuations in their ranking that do not reflect real changes in their performance, this will highlight the limitations of the ranking information and erode confidence in the rankings. As the reputation of the rankings diminishes, the incentive for companies to invest in customer satisfaction weakens. This is particularly a risk for distributing rankings based on quarterly data, which will be subject to even higher levels of noise than the annual rankings.

Report prepared for Ofwat). We have used the scores in Figure 1 for the analysis in this section for two reasons:

(1) Ofwat has used them in the presentation of the incentive scheme; and

(2) The Navigator/Allto data have a smaller sample size than the proposed annual sample size. The resulting scores have a higher standard error and may not represent the spread in scores we would likely see from the proposed design based on annual data.

8 This calculation is based on a pairwise z-test.

9 This probability is estimated by simulation. Please see Annex A for details.

Figure 3 With the current sample size, companies are not statistically different from companies allocated to other reward or penalty buckets

Current sample size

In the payment bands:	Given the set of scores:	**The ranked company:**	**Would be statistically indistinguishable (95% level) from companies with the ranks:**																
Highest payment	84.4	**1 →**	1	2	3	4	5												
Highest payment	83.4	**2 →**	1	2	3	4	5	6	7										
Highest payment	83.2	**3 →**	1	2	3	4	5	6	7										
Highest payment	83	**4 →**	1	2	3	4	5	6	7										
Payment	82.7	**5 →**	1	2	3	4	5	6	7										
Payment	81.1	**6 →**		2	3	4	5	6	7	8	9	10	11						
Payment	80.8	**7 →**		2	3	4	5	6	7	8	9	10	11	12					
No payment	79.8	**8 →**						6	7	8	9	10	11	12	13				
No payment	79.7	**9 →**						6	7	8	9	10	11	12	13				
No payment	79.4	**10 →**						6	7	8	9	10	11	12	13				
No payment	79.3	**11 →**						6	7	8	9	10	11	12	13				
Penalty	78.3	**12 →**							7	8	9	10	11	12	13	14			
Penalty	77.2	**13 →**								8	9	10	11	12	13	14	15	16	
Highest penalty	76.2	**14 →**												12	13	14	15	16	17
Highest penalty	75.5	**15 →**													13	14	15	16	17
Highest penalty	75.1	**16 →**													13	14	15	16	17
Highest penalty	74.1	**17 →**														14	15	16	17

Source: *Frontier Economics; analysis of figures in 'PR19 Customer Measure of Experience (C-MeX): Policy decisions for the C-MeX shadow year 2019-2020' (Confidence interval estimate on p12; distribution of scores on p25)*

Note: *Numbers falling within the diagonal coloured blocks indicate that the two companies are not statistically different and fall within the same payment bucket. Numbers highlighted in dark red indicate that the two companies are not statistically different, but they fall in different payment buckets.*

Figure 4 With the current sample size, there is a high risk of receiving the incorrect rank, and the incorrect payment or penalty

The company assigned to the rank:	Has the risk of having been mis-ranked:
1	48%
2	65%
3	68%
4	69%
5	70%
6	73%
7	74%
8	74%
9	75%
10	74%
11	74%
12	70%
13	66%
14	64%
15	59%
16	56%
17	40%

A company assigned to the payment bucket:	Has the risk of having been allocated to the incorrect payment bucket:
Highest payment	17%
Payment	35%
No payment	36%
Penalty	37%
Highest penalty	11%

Source: *Frontier Economics; analysis of figures in 'PR19 Customer Measure of Experience (C-MeX): Policy decisions for the C-MeX shadow year 2019-2020 (Confidence interval estimate on p12; distribution of scores on p25)*

Figure 4 also shows the risk of misallocating companies to the wrong penalty or reward bucket.[10] For example, there is a 36% chance that a company that has been assigned to the bucket with no payment should have been assigned to a different bucket, if we had been able to observe all companies' scores without noise. This table shows that we cannot be confident that companies end up in the correct penalty or reward bucket. We note several features of this misallocation:

- The buckets are calculated based on the mean and the standard deviation among all company's scores. The noise in one company's score feeds into

[10] These risks are estimated by simulation. Please see Annex A for details.
Figure 4, along with the rest of this section, uses the same distribution of scores displayed in Figure 3.

noise in the mean and standard deviation, which then leads to noise in the location of the cliff-edges. As a result, the noise in one company's score introduces noise in the calculation of the incentive payments or penalties for every other company. The cliff-edges in Ofwat's incentive design have introduced a high cost associated with this misallocation.

- In order to receive the highest possible reward, it is also necessary to rank among the top three companies. As a result, the mis-ranking risks, which are even higher than the misallocation risks, will also contribute to the risk of misallocation for the highest reward category.
- The financial cost of being incorrectly misallocated to a lower performance category will not cancel out the financial benefit of being incorrectly misallocated to a higher performance category in the long run. This is because the rewards from the payment buckets are not equal to the costs in the penalty buckets. For example, consider the case where each company's performance is the same from year to year. Consider a company that would have a score equal to the mean of all companies' scores if it were possible to observe all scores without noise. This company should receive no reward or penalty. If the company were to score 1 standard deviation above the mean due to noise, it would have a reward uplift of 3% of residential retail revenue. However, if it were to score 1 standard deviation below the mean due to noise, it would have a penalty of 6% of residential retail revenue. The noise in the overall scores will have a symmetric distribution,[11] but the rewards and penalties will not be symmetric, and so in the long run this company will accrue a net penalty, purely due to noise.
- The current proposal to pool wastewater score information across Hafren Dyfrdwy and Severn Trent will mean that the noise in their overall scores will be correlated.[12] This will increase the noise in the mean and standard deviation estimates used to calculate reward and penalty buckets, relative to the case where the noise in the Hafren Dyfrdwy and Severn Trent scores were independent and all else equal. This noise in the mean and standard deviation estimates will feed into mis-ranking and misallocating rewards and penalties for all other companies.

We conclude that Ofwat's current sample size is too small given the purpose of the survey. This is especially a concern for companies with scores in the middle of the distribution, whose risk of having received the wrong penalty or reward is particularly high. Figures 3 and 4 together show that the current sample size does not give companies adequate incentives to improve performance, as they cannot be sufficiently confident either about their ranking or about their reward or penalty.

Figure 5 shows how, if the same distribution of scores had come from a sample that was twice or four times larger, the reliability of the rankings would improve.[13] Even for a doubled sample size, more companies have scores that are significantly

[11] This is a statistical property of averages, and the C-MeX score is a weighted average of responses.

[12] We have not modelled this correlation, and it is not included in any of our results.

[13] In this section, for ease of interpretation, we hold the set of scores fixed as we increase the hypothetical sample size. However, if the sample size increased, it is likely that the distribution of overall scores would have a smaller spread (standard deviation). All else equal, a tighter distribution of overall scores decreases the reliability of rankings and reward/penalty buckets. Therefore our approach may underestimate the risk of mis-ranking and mis-allocatiion under a sample size 2 or 4 times larger.

different from the scores assigned to other payment buckets, and this is even more the case for a quadrupled sample size.

Figure 5 Although increasing the sample size would increase the reliability of the rankings, the risk of mis-ranking would remain high even at 4 times the current sample size

2x current sample size

In the payment bands:	Given the set of scores:	**The ranked company:**	**Would be statistically indistinguishable (95% level) from companies with the ranks:**																
Highest payment	84.4	**1 →**	1	2	3	4	5												
	83.4	**2 →**	1	2	3	4	5												
	83.2	**3 →**	1	2	3	4	5												
	83	**4 →**	1	2	3	4	5												
Payment	82.7	**5 →**	1	2	3	4	5	6											
	81.1	**6 →**					5	6	7	8	9	10	11						
	80.8	**7 →**						6	7	8	9	10	11						
No payment	79.8	**8 →**						6	7	8	9	10	11	12					
	79.7	**9 →**						6	7	8	9	10	11	12					
	79.4	**10 →**						6	7	8	9	10	11	12					
	79.3	**11 →**						6	7	8	9	10	11	12					
Penalty	78.3	**12 →**								8	9	10	11	12	13				
	77.2	**13 →**												12	13	14	15		
Highest penalty	76.2	**14 →**													13	14	15	16	
	75.5	**15 →**													13	14	15	16	17
	75.1	**16 →**														14	15	16	17
	74.1	**17 →**															15	16	17

4x current sample size

In the payment bands:	Given the set of scores:	**The ranked company:**	**Would be statistically indistinguishable (95% level) from companies with the ranks:**																
Highest payment	84.4	**1 →**	1	2	3														
	83.4	**2 →**	1	2	3	4	5												
	83.2	**3 →**	1	2	3	4	5												
	83	**4 →**		2	3	4	5												
Payment	82.7	**5 →**		2	3	4	5												
	81.1	**6 →**						6	7	8									
	80.8	**7 →**						6	7	8	9								
No payment	79.8	**8 →**						6	7	8	9	10	11						
	79.7	**9 →**							7	8	9	10	11						
	79.4	**10 →**								8	9	10	11	12					
	79.3	**11 →**								8	9	10	11	12					
Penalty	78.3	**12 →**										10	11	12	13				
	77.2	**13 →**												12	13	14			
Highest penalty	76.2	**14 →**													13	14	15	16	
	75.5	**15 →**														14	15	16	
	75.1	**16 →**														14	15	16	17
	74.1	**17 →**																16	17

Source: *Frontier Economics; analysis of figures in 'PR19 Customer Measure of Experience (C-MeX): Policy decisions for the C-MeX shadow year 2019-2020' (confidence interval estimate on p12; distribution of scores on p25)*

Note: *Sample size assumed to scale up proportionally for all survey components.*

Figure 6 The risk of mis-ranking decreases with a larger sample size

The company with the rank:	Risk of mis-ranking:		
	Current sample size	2x current sample size:	4x current sample size:
1	48%	38%	25%
2	65%	60%	55%
3	68%	62%	56%
4	69%	62%	55%
5	70%	62%	51%
6	73%	60%	42%
7	74%	63%	47%
8	74%	69%	60%
9	75%	67%	60%
10	74%	67%	60%
11	74%	66%	58%
12	70%	56%	37%
13	66%	49%	29%
14	64%	52%	38%
15	59%	52%	43%
16	56%	49%	41%
17	40%	29%	18%

Source: Frontier Economics; analysis of figures in 'PR19 Customer Measure of Experience (C-MeX): Policy decisions for the C-MeX shadow year 2019-2020 (Confidence interval estimate on p12; distribution of scores on p25)

Figure 7 The risk of misallocation between payment/penalty buckets decreases with a larger sample size

A company assigned to the payment bucket:	Risk of misallocation:		
	Current sample size	2x current sample size	4x current sample size
Highest payment	17%	12%	9%
Payment	35%	31%	27%
No payment	36%	24%	15%
Penalty	37%	31%	24%
Highest penalty	11%	7%	5%

Source: Frontier Economics; analysis of figures in 'PR19 Customer Measure of Experience (C-MeX): Policy decisions for the C-MeX shadow year 2019-2020 (Confidence interval estimate on p12; distribution of scores on p25)

Note: Sample size assumed to scale up proportionally for all survey components.

Figure 6 shows how the risk of being assigned the incorrect rank decreases with a sample size that is twice and four times larger. However, even with a quadrupled sample size, the risk of an incorrect rank is over 50% for 8 companies. This shows that, even with a substantially larger sample size, many companies could not be confident that they had received the correct ranking.

Figure 7 shows what the probability of being misallocated to the incorrect payment or penalty bucket would have been if this distribution of scores had come from a sample size twice or four times larger. The risk of misallocation decreases as the sample size increases. With a quadrupled sample size, the risk of a company that was allocated to either the highest payment or highest penalty has a risk of misallocation under 10%, but companies that fall in the middle buckets still have a substantial misallocation risk.

We therefore conclude that Ofwat needs to increase the sample size to reduce the risk of misallocating companies to reward/penalty buckets. However, a bigger sample size does still lead to a substantial risk of misallocation, we have considered other ways of improving C-MeX.

The risk of misallocating penalties and rewards could also be reduced by removing the cliff-edges from the incentive design and moving to a more gradual scale. While this does not address the issue of reporting rankings with a low level of confidence, it reduces the total monetary amount that may be misallocated.

Another option is to limit the impact of noise in one company's score on the noise in another company's bucket allocation. For example, the standard deviation that is used to determine the bucket calculation could be predetermined, based on previous years of C-MeX data. This would reduce the misallocation risk that is due to random variation in the standard deviation, while allowing the mean that centres the bucket calculation to vary flexibly year to year.

3 WHICH COMPONENTS CONTRIBUTE THE MOST TO NOISE IN THE C-MEX SCORE?

In the previous section, we have assessed the statistical validity of C-MeX based on the level of noise or uncertainty around individual scores. In this section, we investigate the sources of the noise around the C-MeX scores further.

C-MeX is a complex measure with many sources of noise

C-MeX is a composite measure that captures customers' satisfaction (those that have been in contact with the company), customers' experience (those that may not been in contact with the company) and customers' net promoter score (NPS) which describes whether customers would recommend the company to their friends and family.

The customer service and experience scores are further broken down into sub-sample groups. For customer service, the sample is split into billing and operations contacts and within each sub-sample, Ofwat distinguishes online and telephone contacts. For the experience scores, the sample is split into telephone and face-to-face surveys.

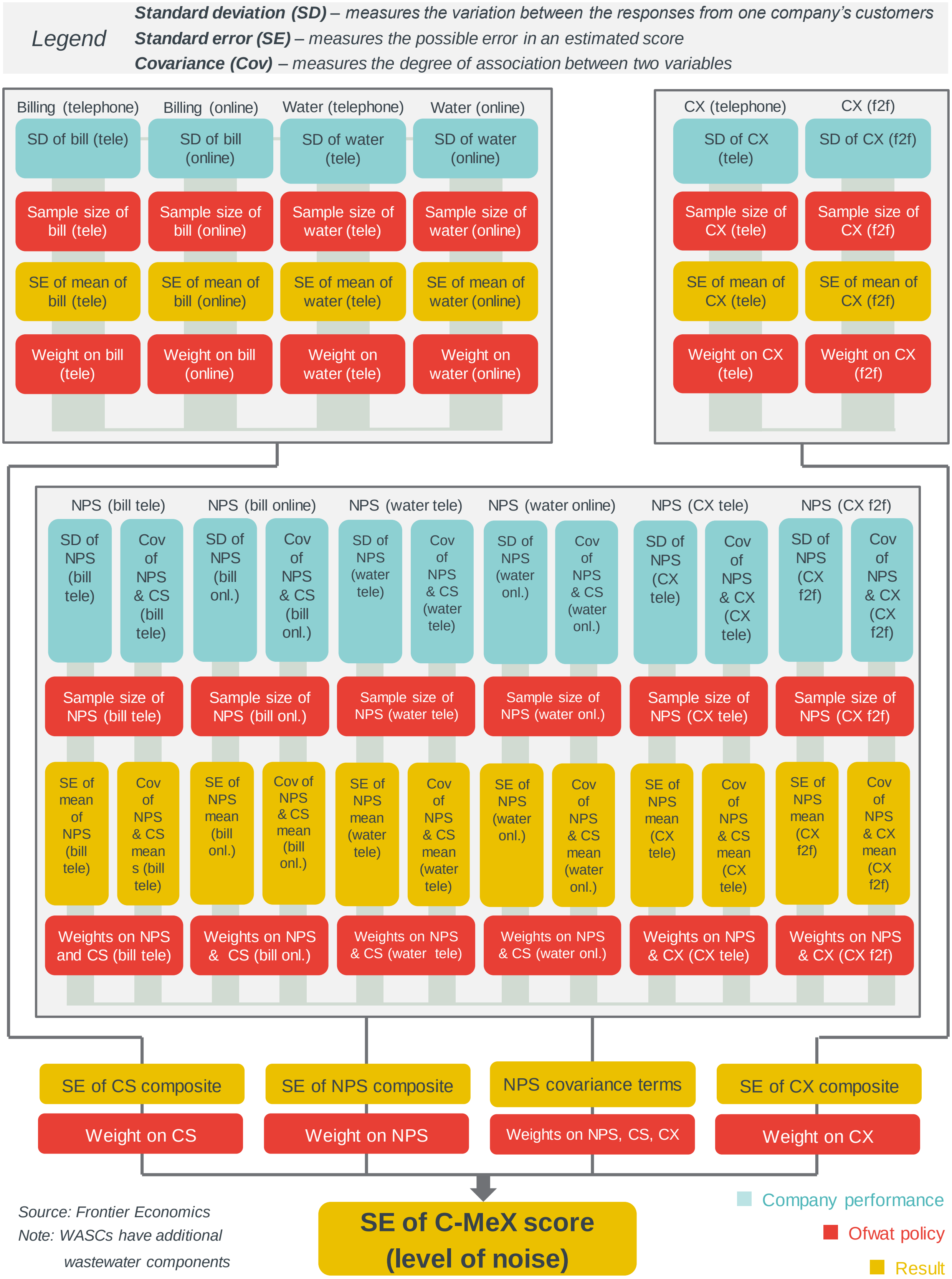
All of the customers are asked about their NPS score but as the question is asked directly after customers have been asked about their customer service or experience there is a high degree of overlap.

Figure 8 below shows all the components that contribute to the level of noise (the standard error) for a company's individual C-MeX score. We have used a WOC calculation in this example, and a WASC would have additional wastewater components. We have colour-coded the sources of noise:

- The blue boxes describe the level of noise in the underlying performance of each company. While changes in company performance will influence the level of noise, this cannot be influenced by the design of C-MeX.
- The red boxes describe Ofwat's policy decisions. These are the policy levers that Ofwat can use to increase or decrease the level of noise in the overall score. Ofwat can determine the sample size of each individual sub-sample and the weightings for each sub-sample.
- The yellow boxes are the components calculated based on the blue and red inputs.

The figure shows that C-MeX is a complex measure and as a result there are many sources of noise. While the complexity is driven by the desire to capture many dimensions of customers' views, it also may impact the overall level of noise in the scores. In the following sections we examine how the different components contribute to the standard error of the overall score.

Figure 8 Components that feed into the C-MeX score standard error (WOC)



Source: Frontier Economics

Note: WASCs have additional wastewater components

Our assessment requires some assumptions that Ofwat could test

Our analysis is based on pilot and shadow year data provided by United Utilities, as we do not have access to each individual company's survey responses. We therefore have to make a number of assumptions:

- The level of noise in the survey responses is stable over time and is the same between companies. There is no *a priori* reason to think that the level of noise will vary significantly between companies. We suggest that Ofwat could replicate our analysis with all of the industry data to test our conclusions.
- All companies achieve Ofwat's targeted sample sizes.
  - We understand that this is not necessarily the case for all sub-samples and is particularly challenging for small companies. If small companies are unable to achieve the target number of surveys within a particular sub-sample, then that sub-sample would be a relatively larger source of noise for that company's overall score.
  - If a company does not achieve the target sample, the company's score will be less precise than if they had managed to achieve the target.
- All WASCs have the same ratio of digital to non-digital contacts, and all WOCs have the same ratio of digital to non-digital contacts. We have used the average ratio among WASCs and among WOCs in the pilot year.

The NPS component contributes significantly to the noise in C-MeX score

Using United Utilities' pilot and shadow year data as an input, we have assessed the level of noise in the overall score.[14] Figure 9 shows the contribution that each component makes to the overall noise in the score.[15] It shows that the experience score contributes around 10% and the customer service score contributes around 25%. The biggest contributor is the NPS score, which includes both the noise in the score itself but also the additional noise resulting from the correlation between NPS and the other measures.[16] This is because customers are asked about their level of customer service or customer experience satisfaction, then their reasons for the score, and then directly afterwards are asked about how likely they are to recommend the company to friends or family. The scores that customers give for satisfaction and NPS are very similar. This is not surprising. Behavioural economics suggests that survey respondents will give internally consistent answers, particularly if they have just been asked to justify a score.

---

[14] United Utilities' survey responses were used to estimate the standard deviation of the responses in every C-MeX sub-component. These standard deviations then fed into the calculation of the standard error of every composite measure, including the overall score. For details of this calculation, please see Annex A.

[15] In Figure 8 we have used the variance as a measure of noise. It captures the same kind of information as the standard error, but it is rescaled to have convenient statistical properties. Unlike standard errors, variances are additive, which means that each C-MeX component contributes a percentage to the total variance of the overall score.

[16] The customer service and experience questions are fielded to different respondents, and so the noise in the customer service measure will be uncorrelated with the noise in the experience measure. However, the same respondent can answer both the customer service and the NPS question, and so the noise in these two measures will be correlated. Similarly, the noise in the customer experience and NPS measure will be correlated. These correlations will contribute to the noise in the overall score. Note that in Figure 7, we present figures using the covariance between NPS and the other components; covariance is another measure of how two variables are associated with one another.

From a sample of UU C-MeX survey interview recordings, there is evidence that many survey responders do not to understand the difference between the satisfaction and NPS questions. A review of customer comments identified multiple examples of customers either providing near identical justification for satisfaction and NPS scores, or directly referencing their previous satisfaction score when justifying their NPS score. For example when commenting on NPS scores customers gave responses such as "*Please see previous comments*"; "*See the first answer!*" and "*The reason I have given for the last question*".

We estimate that dropping the NPS measure from the C-MeX score would decrease the standard error of the C-MeX score by 27% (WASCs and WOCs). This is a drop in standard error equivalent to increasing the sample size by close to 90%. Removing NPS provides substantial benefits in addition to other changes discussed in the previous section.

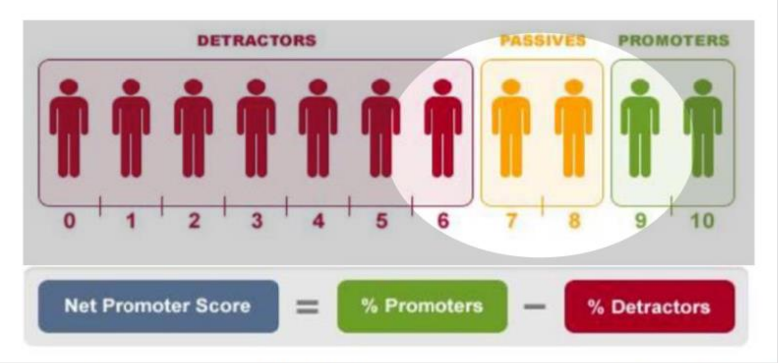
Figure 9 The NPS contributes significantly to the C-MeX score variance

Company type	Component	% contribution to variance of overall score
WASC	CS	24%
	CX	10%
	NPS	23%
	NPS covariance with CS and CX	43%
WOC	CS	25%
	CX	9%
	NPS	22%
	NPS covariance with CS and CX	44%

Source: Frontier Economics; analysis of UU pilot and shadow year data

THE NPS SCORE HAS SUB-OPTIMAL STATISTICAL FEATURES

In addition to being highly correlated with the customer satisfaction and experience scores, the NPS score has sub-optimal statistical features. Survey respondents give an NPS score between 0 and 10. Their response is then converted into the score by using the approach illustrated below. Each answer between 0 and 6 is treated as a score of -100 while each answer between 9 and 10 is treated as +100 to derive the overall score on a 100 point scale. Answers between 7 and 8 are assigned a score of 0. This means that the score is highly sensitive to respondents making selections between 6 and 7 as well as 8 and 9. This approach increases the level of noise in the score.



Source: 'PR19 Customer Measure of Experience (C-MeX): Policy decisions for the C-MeX shadow year 2019-2020'

Customer satisfaction is a noisier measure than customer experience

If the NPS scores were omitted, it is useful to consider the relative contribution of the customer satisfaction and experience measures to the overall level of noise. Figure 10 shows that the satisfaction measure contributes more than twice as much to the noise as the experience measure, based on United Utilities data and Ofwat's target sample sizes.

Ofwat has equal recommended sample sizes for customer service and experience, and equal overall weights on the two measures. However, it is very likely the case that the variation between respondents is not equal for customer service and experience, not only for United Utilities, but also for other companies. A survey question that produces more variable responses will require a larger sample size in order to achieve a particular level of precision for the score on that question. Ofwat might investigate how sample size is allocated between customer service and experience. This analysis could use pilot and shadow year data from the whole industry, and take into account differences between the target sample sizes and the sample sizes that companies were able to achieve. Based on our analysis of one company's data, we would expect Ofwat to find some efficiency gains in the reallocation of survey resources.

We estimate the combined effect of dropping the NPS measure from the C-MeX score and also reallocating Ofwat's proposed annual sample size between CS and CX. We estimate that these two design changes would decrease the standard error of the C-MeX score by 29% (WASCs and WOCs). This is a drop in standard error equivalent to increasing the sample size by nearly 100%.[17] As this estimate is based on United Utilities data, if the variability in United Utilities data differs from that of the whole industry, this may not reflect the efficiency gains of reweighting for all companies.[18] Additionally, it may be sensible to only partially adjust sample sizes based on information about the variability of different components, to maintain reasonably large sample sizes in all components. Ofwat may want to impose minimum sample sizes for components to assure that each is individually measured with sufficient precision.

Figure 10 If the NPS were omitted, the CS and CX measures would not contribute equally to the variance of the overall score

Company type	Measure	% contribution to variance of overall score
WASC	CS	71%
	CX	29%
WOC	CS	73%
	CX	27%

Source: Frontier Economics; analysis of UU pilot and shadow year data

It is possible not only to vary the sample size allocations between CS and CX, but also the sample size allocations between different components within CS and within CX. This may lead to improvements in the precision of the overall score if there are substantial differences in the variability of responses between components, for example between telephone and face-to-face CX respondents. We note that this analysis depends on United Utilities' data at a more granular level, and it may not be representative of other companies.

We estimated that this more granular sample size reallocation, in combination with omitting NPS, would decrease the standard error of the overall score by 33% for a WASC. This is the increase in precision that would be expected from increasing the sample size by 125% with the current C-MeX design. In particular, based on United Utilities' data, it was optimal to reduce the face-to-face CX sample size by around 20%. This figure may not be representative of the whole industry,[19] but it suggests that there may be efficient reallocations of resources. Given that there is sufficient pilot and shadow year data to inform a refining of the survey design, we recommend that Ofwat examine the efficient allocation of sample size. As we noted above, it may be sensible to only partially adjust sample sizes based on information about the variability of different components, to maintain reasonably large sample sizes in all components. For the remainder of this section, we will use Ofwat's recommended sample size allocations within CS and within CX.

[17] We allowed the allocation of the total annual sample size of 1,600 between CS and CX to vary, but assumed the proportions of sample sizes between sub-components within CS, and also within CX, remained constant.

[18] This estimate may not be representative of the whole industry for two reasons. First, United Utilities may have different levels of variability among respondents than the rest of the industry. Second, this calculation uses variances that we have estimated from sample data, and the variance estimates will contain noise.

[19] This estimate has the same limitations that were described in the previous footnote.

We have estimated the impact of our proposed design changes on the risks of mis-ranking and misallocation between reward and penalty buckets.[20] Figure 11 compares the probability of having been mis-ranked, for every company, under three scenarios. The first scenario is the current C-MeX design, the second scenario omits NPS and reallocates sample size between CS and CX, and the third scenario also includes a doubled sample size. This shows that there are moderate improvements in ranking precision that can be achieved without increasing the sample size. However, this improvement is not sufficient to produce reliable rankings: even with design changes and a doubled sample size, 9 companies still have a mis-ranking risk over 50%. As before, this shows that the distribution of scores is too close together to be able to reliably rank.

Figure 12 shows the probability of having been misallocated to the incorrect penalty or reward bucket, under the three scenarios from Figure 11. The design improvements decrease the risk of misallocation substantially; for example, the misallocation risk for the ‘no payment’ bucket decreases from 37% to 28% when dropping NPS and reallocating sample size. Even with this improvement, there is still significant risk of misallocation. This shows that further design changes should be considered, such as removing cliff-edges between reward and penalty buckets.

[20] For this calculation, we have estimated the percentage reduction in the standard error that would result from our proposed design changes, based on UU’s data. We then apply this reduction to the standard error of the overall score that Ofwat has implied. We then use this reduced overall score standard error to estimate the resulting mis-ranking and misallocation risks.

Figure 11 C-MeX design changes can reduce the risk of mis-ranking companies

The company with the rank:	Has the risk of having received the incorrect rank:		
	Current design	Omitting NPS, reallocating sample size	Omitting NPS, reallocating sample size, 2x sample size
1	48%	41%	29%
2	65%	61%	56%
3	68%	64%	58%
4	69%	66%	58%
5	70%	66%	56%
6	73%	67%	49%
7	74%	68%	53%
8	74%	71%	64%
9	75%	71%	65%
10	74%	70%	63%
11	74%	69%	61%
12	70%	62%	44%
13	66%	56%	36%
14	64%	56%	43%
15	59%	55%	47%
16	56%	53%	44%
17	40%	34%	22%

Source: Frontier Economics

Figure 12 C-MeX design changes can reduce the risk of misallocation to the incorrect reward or penalty bucket

A company assigned to the bucket:	Has the risk of misallocation:		
	Current design	Omitting NPS and reallocating sample size	Omitting NPS, reallocating sample size, 2x sample size
Highest payment	17%	13%	10%
Payment	35%	33%	29%
No payment	37%	28%	18%
Penalty	37%	34%	27%
Highest penalty	11%	8%	6%

Source: Frontier Economics

4 KEY RECOMMENDATIONS

We have conducted an independent and objective assessment of the statistical validity of C-MeX. Our findings are:

- The current sample size is not sufficient to be confident in distinguishing the scores. Given Ofwat's current sample size of 1,600, we cannot be confident that there is a real performance difference between companies with annual scores less than 2.7 points apart.
- Companies cannot be confident that their rankings reflect actual differences in company performance. We estimate that a company faces a chance of having received the wrong rank that ranges between 40% and 75% under the proposed design.
- There is a high risk of misallocating rewards and penalties, which is over 10% for all buckets, and is highest for the 'penalty' bucket, at 37% risk of misallocation under the proposed design. The magnitude of these risks is exacerbated by the cliff-edges that are part of the incentive design.
- The risk of mis-ranking and misallocated rewards and penalties is worsened if a company cannot make their sample size targets. Because the rank and payment for each company depends on the score for every company, a company that fails to achieve the recommended sample size increases the risk of mis-ranking and misallocation for every other company.
- Due to the asymmetric distribution of rewards and penalties, a company's payments will not average out to the correct value in the long run.
- Including the NPS score contributes substantially to the noise in the overall scores. NPS accounts for over half of the variance of the overall score, both through its own noise and through its correlation with the CS and CX measures.
- Even if all companies in fact perform very similarly to one another, the proposed design would still generate large differences in payment quantities out of minor differences in customer satisfaction. This would be true even if every company's performance were perfectly observed, without any noise.

Based on our findings, we recommend the following to increase the level of confidence in the overall scores:

- Ofwat should increase the sample size. A larger sample size would substantially reduce the risk that a company has been mis-ranked. The company in the proposed design with the highest mis-ranking risk, at 75%, would see this risk reduced to 67% with a doubled sample size, or 60% with a quadrupled sample size. A larger sample size would also reduce the risk that a company has been allocated to the incorrect reward or penalty bucket. The bucket with the highest risk of misallocation under the proposed design, at 37%, would see this risk reduced to 31% with a doubled sample size, or 24% with a quadrupled sample size.
- Ofwat should remove NPS from the calculation of the overall C-MeX score. We estimate that dropping the NPS measure from the C-MeX score would

decrease the standard error of the C-MeX score by 27%. This is a drop in standard error equivalent to increasing the sample size by close to 90%.

- Ofwat should consider how the sample size is divided between the customer service and experience components to reduce the overall level of noise. We estimate that both omitting NPS and redistributing survey resources from customer experience to customer satisfaction would increase precision by around 30%, which is the improvement one would expect from increasing the sample size by nearly 100%. Ofwat may be able to achieve a higher increase in precision by further considering the allocation of resources within the different sub-components of the surveys.
- Ofwat should also consider removing the “cliff-edges” from the rewards and penalties. This would reduce the total amount of reward and penalty that is misallocated.

To make a final decision on C-MeX, we recommend for Ofwat to replicate our analysis using industry data. This would allow Ofwat to make informed choices about the final sample size and split between components. Annex A describes our calculations in detail.

ANNEX A TECHNICAL ANNEX

This Annex provides details on the following calculations:

- Testing whether two scores are significantly different from one another (Z-test)
- Estimating the probability that a company received the incorrect rank
- Estimating the probability that a company was allocated to the incorrect reward/penalty bucket
- Estimating the contributions of components to the variance of the overall score
- Estimating the optimal allocation of sample size between CS and CX, or between more granular components

A.1.1 Notation and statistical terms

Notation

This section explains notation that will be used in the rest of the Annex.

We use subscripts a, b, c ... to index over companies.

$\bar{Y}_{a,CMEX}$ denotes the observed overall score for company a. $\bar{Y}_{a,CS}, \bar{Y}_{a,CX}, \bar{Y}_{a,NPS}$ denote the CS, CX, and NPS score for company a.

$Y_{a,CMEX}$ denotes the 'true' overall score for company a, which we are trying to estimate with $\bar{Y}_{a,CMEX}$.

$SE_{a,CMEX}$, $\sigma^2_{a,CMEX}$ denote the standard error and variance of the overall score for company a, with analogous notation for the standard error and variance of the score components.

A.1.2 Testing whether two scores are significantly different from one another (Z-test)

Estimate the standard error of the overall score

We can back out the standard error estimate $\widehat{SE}_{CMEX}$ from the confidence interval that Ofwat provided by using the following formula for a 95% confidence interval:

$$CI\ radius = 1.96\, \widehat{SE}_{CMEX}$$

Estimate the minimum significant difference between two scores, with the current sample size

The statistic $Z_{a:b,CMEX}$ for the two-sample test for a difference in normally-distributed means is:

$$Z_{a:b,CMEX} = \frac{\bar{Y}_{a,CMEX} - \bar{Y}_{b,CMEX}}{\sqrt{SE^2_{a,CMEX} + SE^2_{b,CMEX}}} = \frac{\bar{Y}_{a,CMEX} - \bar{Y}_{b,CMEX}}{\sqrt{\frac{\sigma^2_{a,CMEX}}{n_{a,CMEX}} + \frac{\sigma^2_{b,CMEX}}{n_{b,CMEX}}}} \qquad (A.1.2.1)$$

This test will find a significant difference at the 95% level if $Z_{a:b,CMEX} > 1.96$.

We assume that every company's overall score has the same variance and sample size, i.e.

$$\sigma^2_{CMEX} = \sigma^2_{a,CMEX} = \sigma^2_{b,CMEX}, \quad n_{CMEX} = n_{a,CMEX} = n_{b,CMEX}$$

Which means that Equation (A.1.2.1) simplifies to

$$Z_{a:b,CMEX} = \frac{\bar{Y}_{a,CMEX} - \bar{Y}_{b,CMEX}}{\sqrt{2\, SE_{CMEX}}} = \frac{\bar{Y}_{a,CMEX} - \bar{Y}_{b,CMEX}}{\sqrt{\frac{2\,\sigma^2_{CMEX}}{n_{CMEX}}}} \qquad (A.1.2.2)$$

We can rearrange this to solve for the minimum distinguishable difference under the current sample size as a function of our estimate of the standard error $\widehat{SE}_{CMEX}$:

$$\bar{Y}_{a,CMEX} - \bar{Y}_{b,CMEX} = 1.96\sqrt{2\, SE_{CMEX}}$$

Estimate the minimum significant difference between two scores, if the sample size changed

We also can see from Equation (A.1.2.2) that $\bar{Y}_{a,CMEX} - \bar{Y}_{b,CMEX}$ will scale with the sample size by a factor of $1/\sqrt{n_{CMEX}}$.

Suppose we wanted to know the minimum detectible difference $\bar{Y}^*{}_{a,CMEX} - \bar{Y}^*{}_{b,CMEX}$ corresponding to some sample size n^*_{CMEX}. This would be:

$$\bar{Y}^*{}_{a,CMEX} - \bar{Y}^*{}_{b,CMEX} = \sqrt{\frac{n_{CMEX}}{n^*_{CMEX}}}\left(\bar{Y}_{a,CMEX} - \bar{Y}_{b,CMEX}\right)$$

A.1.3 Estimating the probability that a company received the incorrect rank

We run a simulation to estimate the probability of a company receiving an incorrect rank. The simulation looks at what would happen on average if we were able to collect data on all companies many times, each time recalculating the company ranks.

We start with a observed set of 17 company overall scores: $\bar{Y}_{a,CMEX}$, $\bar{Y}_{b,CMEX}$, … $\bar{Y}_{q,CMEX}$. Each of these scores has an associated standard error (in this case we assume all of these standard errors are the same): SE_{CMEX}. Given the data that has been collected, the 'true' score for each company x will be asymptotically normally distributed around the observed score:

$$Y_{x,CMEX} \sim N\left(\mu = \bar{Y}_{x,CMEX}, \sigma = SE_{CMEX}\right), \forall\, x \in \{a, b, \dots, q\}.$$

Iteration 1

Using a normal random number generator, we can draw a 1st set of simulated 'true' scores based on the data; denote these $Y^{(1)}_{a,CMEX}$, $Y^{(1)}_{b,CMEX}$, … $Y^{(1)}_{q,CMEX}$.

Conditional on this simulated ‘true’ score, a simulated observed score for each company x will be asymptotically normally distributed around the simulated ‘true’ score. Based on this simulated ‘true’ score, we can draw a simulated ‘observed’ score. Denote these $\bar{Y}^{(1)}_{a,CMEX}$, $\bar{Y}^{(1)}_{b,CMEX}$, … $\bar{Y}^{(1)}_{q,CMEX}$.

We can now calculate the ranks of the simulated ‘true’ scores (the ranks we would have observed if all scores had no noise), and the ranks of the ‘observed’ scores (the ranks we observe with noisy scores). Every company $a, b, \ldots, q$ can then be classified as correctly or incorrectly ranked in the 1st draw of the simulation.

Repeat

We can then repeat this process of drawing true and observe scores and comparing their ranks many times (we used 10,000 iterations).

Summarise

We then calculate the average mis-ranking rate for every company across the iterations. This is an estimate of the probability of mis-ranking the company.

A.1.4 Estimating the probability that a company was allocated to the incorrect payment/penalty bucket

We estimate the probability of bucket misallocation in the analogous way to the method in the previous section.

Iteration 1

As before, we draw simulated ‘true’ and ‘observed’ scores $Y^{(1)}_{a,CMEX}$, $Y^{(1)}_{b,CMEX}$, … $Y^{(1)}_{q,CMEX}$ and $\bar{Y}^{(1)}_{a,CMEX}$, $\bar{Y}^{(1)}_{b,CMEX}$, … $\bar{Y}^{(1)}_{q,CMEX}$.

We can now calculate the payment/penalty buckets for the ‘true’ scores. This involves calculating the mean and standard deviation of the simulated ‘true’ scores, and assign each a payment/penalty bucket according to Ofwat’s proposed design.

Separately, we assign a payment/penalty bucket to the ‘observed’ scores, also according to Ofwat’s proposed design.

We then calculate whether the bucket assigned to the ‘true’ score was the same as the bucket assigned to the ‘observed’ score, for each company.

For every payment/penalty bucket, from ‘highest payment’ to ‘highest penalty’, we calculate the fraction of companies with that ‘observed’ bucket allocation that had a ‘true’ allocation to a different bucket. We record this misallocation rate.

Repeat

We can then repeat this process of drawing true and observe scores and comparing their assigned buckets many times (we used 10,000 iterations).

Summarise

We then take the average misallocation rate for each bucket across the iterations. This is an estimate of the probability that a company assigned to a particular bucket should have been assigned to a different bucket.

A.1.5 Estimating how components contribute to the variance of the overall score

In this section, we show how to turn the calculation of the C-MeX score into a calculation of the C-MeX variance for a particular company. We will calculate the variance for some arbitrary target sample size that may not correspond to the sample size of the collected data.

Notation and definitions

As we only consider one company, we omit the subscript for a particular company. Instead, the subscript s in this section indexes over survey components. For NPS components, we denote these with a s, NPS subscript, where s is the CS or CX component that the same respondent will have answered. An example of this would be: the customer service responses for billing (by telephone), and the NPS responses given by those surveyed for billing (by telephone).

Let $Y_{s,i}$ be the survey response for person i on component s. Let σ_s^2 refer to the variance among respondents within the component, let n_s^{obs} refer to the number of observed respondents for the component, and let n_s^{target} refer to the target number of respondents for the component.

Let w_s refer to the weight on the component within the measure. For example, face-to-face contacts have a weight of 0.2 within the CX measure. The sum of all w_s for CS components will be 1; as will the sum of all w_s for CX components; as will the sum of all w_s for NPS components.

Properties of a variance

The calculation below uses the following properties of variances.

First, the weighted average of two **independent** random variables X and Y is a weighted average of the variance, but where the weights have been squared. This is due to the fact that the variances have squared units:

$$\mathrm{Var}(aX + bY) = \mathrm{a}^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)\,, \quad X \textit{ and } Y \textit{ independent}$$

Second, the weighted average of two **dependent** random variables will need to take into account that the variables co-move. So it is a weighted average of their variances and the covariance:

$$\mathrm{Var}(aX + bY) = \mathrm{a}^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y)$$

Rescale the C-MeX components

The variance calculation below requires that a company's overall score is a weighted average of observations. First it is necessary to rescale the scores so that this is the case.

We have performed all of these calculations so that the scale of the overall score is 0-100. This involves multiplying CS and CX responses by 10 so that they are on a scale of 0-100. It also involves recoding NPS scores so that responses from 0 to 6 are recoded -100, 7 and 8 are recoded 0, 9 and 10 are recoded 100.

Estimate variances and covariances of components

The sample variance for a particular component, $\hat{\sigma}_s^2$ is essentially the average squared deviation of an observation from its mean (i.e. the square of the standard deviation) across all respondents i for that component:

$$\hat{\sigma}_s^2 = \frac{\sum_{i=1}^{n_s^{obs}} \left(Y_{s,i} - \bar{Y}_s\right)^2}{n_s^{obs} - 1}$$

Every pair of questions that were asked of single respondents will have a covariance. In other words, there will be an NPS covariance term for every CS and CX component. The sample covariance is essentially the average product of the deviations of two variables from their means:

$$\widehat{Cov}_{s,NPS}^2 = \frac{\sum_{i=1}^{n_s^{obs}} \left(Y_{s,i} - \bar{Y}_s\right)\left(Y_{s,NPS,i} - \bar{Y}_{s,NPS}\right)}{n_s^{obs} - 1}$$

Note that n_s^{obs} should be adjusted if respondents failed to answer both the CS/CX question and the NPS question.

The variances and covariances of the measures

The estimated variances of CS, CX, and NPS are the sums of their respective component variances, rescaled by the squared weight and sample size:

$$\hat{\sigma}_{CS}^2 = \sum_{s \in CS} \frac{w_s^2\, \hat{\sigma}_s^2}{n_s^{target}}$$

The combined contribution of the NPS covariance terms is:

$$\widehat{Cov}_{NPS} = \sum_{s \in CS\ or\ CX} \frac{w_s\, w_{s,NPS}\, \widehat{Cov}_{s,NPS}^2}{n_s^{target}}$$

Contributions to the overall variance

The variance of the overall score depends on the CS 40%/ CX 40%/ NPS 20% weight split:

$$\hat{\sigma}_{CMEX}^2 = 0.4^2\, \hat{\sigma}_{CS}^2 + 0.4^2\, \hat{\sigma}_{CX}^2 + 0.2^2\, \hat{\sigma}_{NPS}^2 + 2 * (0.4 * 0.2)\, \widehat{Cov}_{NPS}$$

So the contributions of each component to the variance of the overall score are:

$$\frac{0.4^2\ \hat{\sigma}_{CS}^2}{\hat{\sigma}_{CMEX}^2}, \frac{0.4^2\ \hat{\sigma}_{CX}^2}{\hat{\sigma}_{CMEX}^2}, \frac{0.4^2\ \hat{\sigma}_{NPS}^2}{\hat{\sigma}_{CMEX}^2}, \frac{2 * (0.4 * 0.2)\ \widehat{Cov}_{NPS}}{\hat{\sigma}_{CMEX}^2}$$

A.1.6 Estimating the optimal allocation of sample size between CS and CX, or between granular components

Supposing that NPS were omitted from C-MeX, that CS and CX were each weighted 50% in the overall score, and the current total sample size n_{CMEX} were held fixed, we can estimate how the sample might optimally be allocated between CS and CX. As the CS and CX scores are weighted averages of their components, their variances will scale by $1/n$.

Let $n_{CS}, n_{CX} = n_{CMEX} - n_{CS}$ be the current sample sizes. Let $\sigma_{CS}^2, \sigma_{CX}^2$ be the variances of the CS and CX scores under the current sample sizes. The optimal allocation $n_{CS}^*, n_{CX}^* = n_{CMEX} - n_{CS}^*$ between CS and CX will minimise the variance of the overall score:

$$n_{CS}^*, n_{CX}^* = \underset{n_{CS}^*, n_{CX}^* > 0}{\arg\min} \left(\frac{w_{CS}^2 \sigma_{CS}^2 n_{CS}}{n_{CS}^*} + \frac{w_{CX}^2 \sigma_{CX}^2 n_{CX}}{n_{CX}^*} \right)$$

$$\text{such that} \quad n_{CS}^* + n_{CX}^* = n$$

The degree to which the optimal sample size should be scaled up or down, n_{CS}^*/n_{CS}, is given by the solution:

$$\frac{n_{CS}^*}{n_{CS}} = \frac{w_{CX}^{-1}\sigma_{CX}^{-1}}{w_{CS}^{-1}\sigma_{CS}^{-1} + w_{CX}^{-1}\sigma_{CX}^{-1}}\, n_{CMEX}, \quad \frac{n_{CX}^*}{n_{CX}} = \frac{w_{CS}^{-1}\sigma_{CS}^{-1}}{w_{CS}^{-1}\sigma_{CS}^{-1} + w_{CX}^{-1}\sigma_{CX}^{-1}}\, n_{CMEX}$$

Intuitively, this means that the optimal sample sizes for CS and CX are determined by the relative variability of the two populations. This generalises to allocating sample size to more than two components. Suppose the components of C-MeX are indexed over s, as in the previous section, and we would like to estimate the optimal $n_1^*, n_2^*, \dots, n_S^*$ across components with variances among respondents $\sigma_1^2, \sigma_2^2, \dots, \sigma_S^2$.

$$n_1^*, n_2^*, \dots, n_S^* = \underset{n_1^*, n_2^*, \dots, n_S^* > 0}{\arg\min} \left(\frac{w_1^2 \sigma_1^2}{n_1^*} + \frac{w_2^2 \sigma_2^2}{n_2^*} + \dots + \frac{w_S^2 \sigma_S^2}{n_S^*} \right)$$

$$\text{such that} \quad n_1^* + n_2^* + \dots + n_S^* = n$$

We can see that at the optimum, the gradient of each term with respect to n_s^* will be equalised:

$$\frac{w_1^2\sigma_1^2}{(n_1^*)^2} = \frac{w_2^2\sigma_2^2}{(n_2^*)^2} = \dots = \frac{w_S^2\sigma_S^2}{(n_S^*)^2}$$

Intuitively, this means that the precision gains of increasing the sample size will be equalised across all components at the optimum. Suppose we want to solve for n_1^*. We can then plug these equalities back into the constraint:

$$n_1^* + \frac{w_2 \sigma_2 n_1^*}{w_1 \sigma_1} + \dots + \frac{w_S \sigma_S n_1^*}{w_1 \sigma_1} = n$$

Which rearranges to:

$$n_1^* = \frac{w_1 \sigma_1}{\sum_{s=1}^{S} w_s \sigma_s} n$$

In other words, the share of the sample size for component s is proportional to the share of its weighted standard deviation among all components. Intuitively, this means that one should apportion a larger sample size for those components that are more variable and a smaller sample size for those that are less variable.