# Rail Passenger Customer Experience Survey – Method Review

## Overall technical report for the method review

13th November 2023

# Table of contents

# 1 Background

In summer 2022, the rail industry commissioned a method review. The project was managed by the rail industry comprising members of:

- Department for Transport Rail (DfT),
- The Rail Delivery Group (RDG),
- Transport Focus (TF),
- Network Rail (NR),
- Great British Rail Transition Team,

The method review has evaluated a wide range of potential methodologies to use in this enhanced customer satisfaction programme. There were six phases to the method review:

- Phase 1 – development of criteria by which to evaluate different methodologies.
- Phase 2 – development of a weighting process to aggregate scores from the different criteria.
- Phase 3 – analysis of existing documentation, leading to a listing of all possible methodologies.
- Phase 4 – scoring of all possible methodologies against all the criteria agreed in phase 1 and the production of an aggregate score for each methodology using the weightings derived in phase 2.
- Phase 5 – field trial results of the top methodologies and identification of the optimal approach.
- Phase 6 – final recommended (optimal) detailed methodology, confirming how it will meet all the survey requirements and the needs of different users.

For each phase, a technical report was produced which documented the processes that were used in that phase and confirmed the major conclusions that have been derived and agreed.

Now that the method review is complete, these reports have been aggregated into a complete Technical Report which will be peer reviewed by an independent assessor. The Method Review and Technical Reports will together form the specification for the enhanced approach to measuring rail customer satisfaction.

# 2 Outline of phase 1 process

This part of the report lays out the chronology of the process of deriving a set of criteria against which to evaluate a wide range of methodologies. In doing this it is important that the criteria:

- Are comprehensive in covering all key issues which might affect how a methodology works.
- Are succinct, to allow a wide range of individuals to rate the criteria for importance (phase 2).
- Have variance between methods (there is no point rating a wide range of methods on a criterion where the importance is likely to be the same for each method).
- Meet all the essential criteria and objectives for the study as detailed in the minimum viable product (MVP) as detailed in the Statement of Requirements (SOR) for the method review.
- Achieve agreement among the various interests represented in the Technical Group.

Most of the liaison between the agency and the client has been achieved through phases 1-4 via a weekly progress meeting made up of the following individuals:

Jenny Dickson, Rebecca Harpley - Rail Statistics, Research and Evaluation, DfT.

David Greeno - Transport Focus.

Tim Sander, Thomas Folque, Alice Wells – BVA BDRC.

David Chilvers – David Chilvers Associates.

At certain key stages, the whole Technical Working Group has been involved to ensure the widest range of input.

## 2.1 Process of developing the scoring criteria

The scoring criteria were developed through an iterative process of seven stages of feedback on the criteria for assessing methodologies. An initial list of 10 criteria was generated by BVA BDRC on 25th August 2022 and discussed at the weekly meeting on 30th August (please see table 1).

**Table 1: Initial list of scoring Criteria 25th August 2022**

| Criteria | Definition |
|---|---|
| ***1. Coverage of the required universe*** | How well does the method cover the definition of rail users agreed for the survey (this might be those who have travelled in the last week, those who travelled yesterday, those who travelled yesterday/today etc.)? |
| ***2. Ability to generate a random sample*** | How well can the method generate large random samples of rail users? |
| ***3. Ability to generate required weekly sample size*** | How well can the overall target sample size each week be delivered by the method? |
| ***4. Ability to generate required sample sizes of key subgroups*** | How well can the method reach targets for key subgroups such as regions, TOCs and TOC building blocks? |
| ***5. Accuracy of information about the journey undertaken*** | How accurate will data be about the journey being assessed including date, time, stations used and TOCs used? |
| ***6. Incidence rate*** | What percentage of respondents approached are likely to meet the rail use criteria (used in last 7 days/used today/yesterday/used today)? |
| ***7. Response rates*** | What response rate is the method likely to generate? |
| ***8. Speed of generating topline results*** | How quickly can data be produced once the respondent has been contacted using the method? |
| ***9. Weighting*** | How much weighting is likely to be needed to generate results that are representative of journeys undertaken? |
| ***10. Cost*** | What is the estimated cost per complete interview using the method? |

The criteria were revised following this and six further rounds of feedback between 4th September and 27th September 2022.

The feedback identified a range of issues that the criteria had to take into account.

- What definitions used for specifying the universe to be sampled, including whether this should be terms of journeys, rail users and different definitions of rail users depending on the refence journey used. It should also include how different methods capture different stages of the journey.
- Ensuing the scalability of methods, including of subgroups and also how maximum samples were considered. There was also feedback on need for methods to address aggregation of samples over time periods, delivery of independent samples for time periods and also replicability and combining of samples.
- How criteria were to be scored and ranked and inclusion of qualitative assessment of scoring outcomes.
- Potential circularities between criteria or duplication of criteria, and where criteria were in the minimum viable product of the statement of works had not been included.
- Consideration of design effects including weighting efficiency of key subgroups and addressing response bias.
- Inclusion of questionnaire design factors, such as modular approaches.
- Respondents' engagement with the method including fatigue.
- Speed of reporting, how disruption will be dealt with and how costs were considered.

## 2.2 Final criteria

Feedback was considered and included amendments made if they hadn't been included elsewhere by other criteria. The final criteria was agreed and is outlined in table 2 below.

**Table 2: Final scoring criteria 27th September 2022**

| # | Criteria | Definition |
|---|---|---|
| 1 | ***Coverage of the required universe*** | How well does the method cover the definition of rail journeys agreed for the survey (this might be journeys in the last week, journeys yesterday, journeys yesterday/today etc.)? |
| 2 | ***Ability to generate a random representative sample*** | How well can the method generate random samples of rail journeys and how replicable is that? Is there potential for bias or exclusion of key subgroups? How inclusive is the method for specific subgroups e.g., with disabilities? How easy is it to merge data from different methods if a mixed methodology is used? |
| 3 | ***Ability to generate large samples (scalability)*** | How scalable is the method to larger sample sizes overall and to sample sizes for key subgroups? Can this be done without reinterviewing individuals too often? |
| 4 | ***Ability to generate required sample sizes of key subgroups*** | How well can the method reach targets or maximum sample sizes for key subgroups such as regions, TOCs, TOC building blocks and individual stations and where necessary increase sample sizes for key subgroups? |
| 5 | ***Knowing the exact train the person was travelling on*** | How good is the method at being able to allocate the respondent to a specific train with high accuracy? |
| 6 | ***Accuracy of other information about the journey undertaken*** | How accurate will other data be about the journey being assessed including facilities at the station and on the train, other legs of the journey not being monitored, methods of travel to and from stations etc? (We assume that we are still monitoring just one leg of a journey otherwise assignment of TOCs, stations and start times become difficult)? Could the method be open to bias if question responses cannot be randomised or rotated? |
| 7 | ***Incidence rate*** | What percentage of respondents approached are likely to meet the rail use criteria (journey in last 7 days/used today/yesterday)? |
| 8 | ***Response rate*** | What response rate is the method likely to generate? |
| 9 | ***Speed of generating topline results*** | How quickly can data be produced once the respondent has been contacted using the method both overall and at TOC level? |
| 10 | ***Weighting efficiency*** | What is the expected weighting efficiency generated by the method and how does this vary by key subgroups? Does the method reduce weighting efficiency due to any form of clustering e.g., on train or at station? |
| 11 | ***Cost*** | What is the estimated cost per complete interview using the method including all costs specific to the method? |

| # | Criteria | Definition |
|---|---|---|
| 12 | ***Cutting data by different time periods*** | How easy is it to focus on specific time periods and also facilitate aggregation of properly weighted data for different time periods, different subgroups and support comparison of independent samples? |
| 13 | ***Interview length*** | How capable is the method of dealing with longer questionnaires (primarily in terms of respondent fatigue and engagement and also likely dropout rates)? Can the method easily use a modular approach which still enables properly weighted data to be produced for specific time periods? |
| 14 | ***Ability to recontact participants*** | How easy is it to recontact respondents for further information or other projects (i.e., to what extent can contact details be obtained using the method)? |
| 15 | ***The ability to merge with other data*** | How easy is it to merge the CX data with information from other sources (e.g., train data which might show any actual delay of the train or station data which might hold information about the facilities at the station)? |
| 16 | ***Practicability/Feasibility*** | How easy is it to use the methodology in practice? Any there any logistical constraints in using this approach? Can the method cover all the metrics in the questionnaire e.g., to rate the overall end-to-end journey? Is the method susceptible to problems if the rail system is disrupted? |

The final set of criteria that was sent out for evaluation is shown in the pro forma in the next section.

# 3 Outline of phase 2 process

Phase 1 of the method review developed a series of criteria against which each potential methodology will be evaluated. To enable comparisons to be made, a weighting system has been developed which asks the members of the Rail Technical Working Group to score each of the criteria. The proforma based on the final criteria that was distributed is shown in table 3 below.

**Table 3: Final scoring criteria in proforma format**

Rail Customer Experience Survey - method review

| | |
|---|---|
| Please assign a score to each of the criteria in the **Yellow cells in Col C** which reflect how important you think each one is in a potential methodology | |
| The scores can be any numbers you like, they don't have to add up to any particular total. | |
| We have started each score at 10 so you can vary upwards or downwards if you think a particular criterion is of more or less importance | |
| Your scores have been converted to percentages in Col D so you can see these and if necessary modify any of your scores | |
| When you have finalised your scores, please email this Excel sheet to **david.chilvers@dcaweb.co.uk** | |
| Please complete the fields below with your name and email address, so that we can contact you if we have any questions | |

**Name:** **Email:**

| # | Criteria | Definition | Score | Deduced weighting |
|---|---|---|---|---|
| 1 | ***Coverage of the required universe*** | How well does the method cover the definition of rail journeys agreed for the survey (this might be journeys in the last week, journeys yesterday, journeys yesterday/today etc.)? | 10 | 6.3% |
| 2 | ***Ability to generate a random representative sample*** | How well can the method generate random samples of rail journeys and how replicable is that? Is there potential for bias or exclusion of key subgroups? How inclusive is the method | 10 | 6.3% |

| # | Criteria | Definition | Score | Deduced weighting |
|---|---|---|---|---|
| | | for specific subgroups e.g. with disabilities. How easy is it to generate a representative sample by merging data from different methods if a mixed methodology is used? | | |
| 3 | ***Ability to generate large samples (scalability)*** | How scalable is the method to larger sample sizes overall and to sample sizes for key subgroups. Can this be done without reinterviewing individuals too often? | 10 | 6.3% |
| 4 | ***Ability to generate required sample sizes of key subgroups*** | How well can the method reach targets or maximum sample sizes for key subgroups such as regions, TOCs, TOC building blocks and individual stations and where necessary increase sample sizes for key subgroups? | 10 | 6.3% |
| 5 | ***Knowing the exact train the person was travelling on*** | How good is the method at being able to allocate the respondent to a specific train with high accuracy? | 10 | 6.3% |
| 6 | ***Accuracy of other information about the journey undertaken*** | How accurate will other data be about the journey being assessed including facilities at the station and on the train, other legs of the journey not being monitored, methods of travel to and from stations etc? (We assume that we are still monitoring just one leg of a journey otherwise assignment of TOCs, stations and start times become difficult) Could the method be open to bias if question responses cannot be randomised or rotated? | 10 | 6.3% |
| 7 | ***Incidence rate*** | What percentage of respondents approached are likely to meet the rail use criteria (journey in last 7 days/used today/yesterday)? | 10 | 6.3% |
| 8 | ***Response rate*** | What response rate is the method likely to generate? | 10 | 6.3% |
| 9 | ***Speed of generating topline results*** | How quickly can data be produced once the respondent has been contacted using the method both overall and at TOC level? | 10 | 6.3% |
| 10 | ***Weighting efficiency*** | What is the expected weighting efficiency generated by the method and how does this vary by key subgroups? Does the method reduce weighting efficiency due to any form of clustering e.g. on train or at station? | 10 | 6.3% |
| 11 | ***Cost*** | What is the estimated cost per complete interview using the method including all costs specific to the method? | 10 | 6.3% |
| 12 | ***Cutting data by different time periods*** | How easy is it to focus on specific time periods and also facilitate aggregation of properly weighted data for different time periods, different subgroups and support comparison of independent samples? | 10 | 6.3% |
| 13 | ***Interview length*** | How capable is the method of dealing with longer questionnaires (primarily in terms of respondent fatigue and engagement and also likely dropout rates). Can the method easily | 10 | 6.3% |

| # | Criteria | Definition | Score | Deduced weighting |
|---|---|---|---|---|
| | | use a modular approach which still enables properly weighted data to be produced for specific time periods? | | |
| 14 | ***Ability to recontact participants*** | How easy is it to recontact respondents for further information or other projects (i.e. to what extent can contact details be obtained using the method)? | 10 | 6.3% |
| 15 | ***The ability to merge with other data*** | How easy is it to merge the CX data with information from other sources (e.g. train data which might show any actual delay of the train or station data which might hold information about the facilities at the station)? | 10 | 6.3% |
| 16 | ***Practicability/Feasibility*** | How easy is it to use the methodology in practice? Any there any logistical constraints in using this approach? Can the method cover all the metrics in the questionnaire e.g. to rate the overall end-to-end journey? Is the method susceptible to problems if the rail system is disrupted? | 10 | 6.3% |
| | | | | |
| | | Total | 160 | 100% |

## 3.1 Individuals providing feedback

The proforma was sent out to the nine members of the Technical Working Group on 29th September 2022 (and DfT shared this with another colleague also). The individuals on the Technical Working Group, which includes research and insight professionals from each organisation, are outlined in table 4.

**Table 4: Representatives and organisations on technical Working Group 29th September 2022**

| Name | Organisation |
|---|---|
| Jenny Dickson | Department for Transport |
| Rebecca Harpley | Department for Transport |
| Sheila Honey* | Department for Transport |
| Andrew Regan | Network Rail |
| Amy Slynes | Rail Delivery Group |
| Jason Webb | Rail Delivery Group |
| Trevor Taylor | Rail Delivery Group |
| Ian Wright | Transport Focus |
| Louise Coward | Transport Focus |
| David Greeno | Transport Focus |

*member of procurement group.

The proformas had an initial score of 10 for each of the 16 criteria. Individuals then increased or decreased these in terms of whether they thought a particular criterion was more or less important than others. A further column showed how the current scores convert into percentages to help the individual finetune their scores.

All proformas containing scores were sent out on 30th September 2022 and were received back by 4th October. Most organisations provided an overall score, whilst Transport Focus provided individual scores for each person. All analysis was then undertaken using the average for each organisation for each of the 16 criteria.

## 3.2 Analysis of the scores

A number of metrics were computed from the individual scores. Firstly, the average scores for each of the criteria for each organisation was computed. Then a number of summary statistics were produced for each of the criteria across the four organisations, including:

- The arithmetic mean across the four organisations.
- The median across the four organisations.
- The arithmetic mean for each of the four organisations.

The results of these calculations are shown in table 5 below.

**Table 5: Calculated metrics for each criteria**

| # | Criteria | mean | median | NR | RDG | DfT | TF |
|---|---|---|---|---|---|---|---|
| 1 | *Coverage of the required universe* | 9.2% | 9.2% | 8.6% | 9.5% | 8.8% | 10.0% |
| 2 | *Ability to generate a random representative sample* | 8.6% | 8.7% | 8.6% | 7.1% | 8.8% | 10.0% |
| 3 | *Ability to generate large samples (scalability)* | 8.6% | 8.6% | 8.6% | 9.5% | 7.5% | 8.7% |
| 4 | *Ability to generate required sample sizes of key subgroups* | 8.9% | 8.8% | 8.6% | 9.5% | 8.8% | 8.7% |
| 5 | *Knowing the exact train the person was travelling on* | 5.5% | 6.1% | 2.6% | 7.1% | 6.1% | 6.1% |
| 6 | *Accuracy of other information about the journey undertaken* | 4.8% | 4.7% | 2.6% | 7.1% | 4.1% | 5.4% |
| 7 | *Incidence rate* | 4.6% | 4.5% | 4.3% | 4.8% | 5.4% | 3.8% |
| 8 | *Response rate* | 5.7% | 5.2% | 5.2% | 4.8% | 7.5% | 5.2% |
| 9 | *Speed of generating topline results* | 5.1% | 5.1% | 6.0% | 4.8% | 5.4% | 4.2% |
| 10 | *Weighting efficiency* | 5.9% | 5.6% | 6.0% | 4.8% | 7.5% | 5.2% |
| 11 | *Cost* | 5.3% | 5.4% | 6.0% | 2.4% | 4.8% | 8.0% |
| 12 | *Cutting data by different time periods* | 6.4% | 6.6% | 6.0% | 7.1% | 7.5% | 4.8% |
| 13 | *Interview length* | 5.1% | 4.8% | 6.0% | 4.8% | 4.8% | 4.8% |
| 14 | *Ability to recontact participants* | 3.8% | 2.9% | 8.6% | 2.4% | 0.7% | 3.4% |
| 15 | *The ability to merge with other data* | 6.0% | 6.4% | 6.0% | 7.1% | 6.8% | 3.8% |
| 16 | *Practicability/Feasibility* | 6.6% | 6.6% | 6.0% | 7.1% | 5.4% | 8.0% |

For some of the criteria, the range of scores was low. For others, some organisations scored differently from the rest e.g., RDG scored much lower on cost than the others and higher on the ability to generate large samples and sample sizes for key subgroups. The outliers identified are identified in table 6 below. The columns in the table identify different metrics,

- "Outlier score" is the score of the outlier organisation for that specific criterion.
- "Min" is the minimum of the other three organisations.
- "Max" is the maximum of the other three organisations.

**Table 6: Metric outliers for individual organisations with indicators of range**

| outlier | # | Criteria | Outlier score | Min | Max | NR | RDG | DfT | TF |
|---|---|---|---|---|---|---|---|---|---|
| DfT | 8 | Response rate | 7.5% | 4.8% | 5.2% | 5.2% | 4.8% | | 5.2% |
| DfT | 10 | Weighting efficiency | 7.5% | 4.8% | 6.0% | 6.0% | 4.8% | | 5.2% |
| outlier | # | Criteria | Outlier score | Min | Max | NR | RDG | DfT | TF |
| NR | 5 | Knowing the exact train the person was travelling on | 2.6% | 6.1% | 7.1% | | 7.1% | 6.1% | 6.1% |
| NR | 6 | Accuracy of other information about the journey undertaken | 2.6% | 4.1% | 7.1% | | 7.1% | 4.1% | 5.4% |
| NR | 13 | Interview length | 6.0% | 4.8% | 4.8% | | 4.8% | 4.8% | 4.8% |
| NR | 14 | Ability to recontact participants | 8.6% | 0.7% | 3.4% | | 2.4% | 0.7% | 3.4% |
| outlier | # | Criteria | Outlier score | Min | Max | NR | RDG | DfT | TF |
| RDG | 2 | Ability to generate a random representative sample | 7.1% | 8.6% | 10.0% | 8.6% | | 8.8% | 10.0% |
| RDG | 3 | Ability to generate large samples (scalability) | 9.5% | 7.5% | 8.7% | 8.6% | | 7.5% | 8.7% |
| RDG | 4 | Ability to generate required sample sizes of key subgroups | 9.5% | 8.6% | 8.8% | 8.6% | | 8.8% | 8.7% |
| RDG | 11 | Cost | 2.4% | 4.8% | 8.0% | 6.0% | | 4.8% | 8.0% |
| outlier | # | Criteria | Outlier score | Min | Max | NR | RDG | DfT | TF |
| TF | 11 | Cost | 8.0% | 2.4% | 6.0% | 6.0% | 2.4% | 4.8% | |
| TF | 12 | Cutting data by different time periods | 4.8% | 6.0% | 7.5% | 6.0% | 7.1% | 7.5% | |
| TF | 15 | The ability to merge with other data | 3.8% | 6.0% | 7.1% | 6.0% | 7.1% | 6.8% | |
| TF | 16 | Practicability/Feasibility | 8.0% | 5.4% | 7.1% | 6.0% | 7.1% | 5.4% | |

We went back to each organisation detailing where their scores were outliers (either high or low) to seek feedback on why this was the case. Each organisation was provided with the above data where their organisation appeared to be an outlier on 5th October and was asked to provide a rationale for why their score appeared to be an outlier.

The feedback identified that weighting efficiency and response rates were outliers for some organisations due to their importance on the reliability and impact on robustness of different granularity of analysis that is required.

Questions were raised about the achievability of accurate recall of journey by respondents so had downrated the importance of this element but increased as the importance of interview length on response rate and engagement as well as the facility for recontacting respondents to do further research and increase the value of the survey.

## 3.3 Final criteria

We have assessed each line of feedback and aggregated into a combined set in table 7 below.

**Table 7: Final criteria with range of scores excluding outliers, and outliers identified with associated feedback**

| # | Criteria | Min | Max | Outlier | Outlier score | Feedback |
|---|---|---|---|---|---|---|
| 2 | Ability to generate a random representative sample | 8.60% | 10.00% | RDG | 7.10% | RDG scored lower on quality than quantity |
| 3 | Ability to generate large samples (scalability) | 7.50% | 8.70% | RDG | 9.50% | RDG scored higher on quantity than quality |
| 4 | Ability to generate required sample sizes of key subgroups | 8.60% | 8.80% | RDG | 9.50% | RDG scored higher on quantity than quality |
| 5 | Knowing the exact train the person was travelling on | 6.10% | 7.10% | NR | 2.60% | The methods will differ on this so possibly push average up a bit to minimise impact of this outlier |
| 6 | Accuracy of other information about the journey undertaken | 4.10% | 7.10% | NR | 2.60% | The methods will differ on this so possibly push average up a bit to minimise impact of this outlier |
| 8 | Response rate | 4.80% | 5.20% | DfT | 7.50% | Push average up a little |
| 10 | Weighting efficiency | 4.80% | 6.00% | DfT | 7.50% | Push average up a little |
| 11 | Cost | 4.80% | 8.00% | RDG | 2.40% | Ignore RDG score |
| 11 | Cost | 2.40% | 6.00% | TF | 8.00% | Push average up a little |
| 12 | Cutting data by different time periods | 6.00% | 7.50% | TF | 4.80% | Push average down a little |
| 13 | Interview length | 4.80% | 4.80% | NR | 6.00% | Other three have same weight so set at this average level |
| 14 | Ability to recontact participants | 0.70% | 3.40% | NR | 8.60% | This is not part of MVP so perhaps reduce average to limit impact of this outlier |
| 15 | The ability to merge with other data | 6.00% | 7.10% | TF | 3.80% | Push average down a little |
| 16 | Practicability/Feasibility | 5.40% | 7.10% | TF | 8.00% | Push average up a little |

The feedback has then been incorporated into an initial weighting score.

## 3.4 Initial weights provided

The feedback was used along with the data to recommend using either the arithmetic mean, the median or some other metric. Medians were used when there was one outlier but sometimes an organisations score has not been included in creating the initial weights.

**Table 8: Scoring weights with calculation basis, rebased to sum to 100%**

| # | Criteria | Initial recommendation | basis | Initial recommendation (rebased to 100%) |
|---|---|---|---|---|
| 1 | *Coverage of the required universe* | 9.2% | mean | 9.0% |
| 2 | *Ability to generate a random representative sample* | 9.1% | mean excluding RDG | 8.9% |
| 3 | *Ability to generate large samples (scalability)* | 8.6% | median | 8.4% |
| 4 | *Ability to generate required sample sizes of key subgroups* | 8.9% | median | 8.7% |
| 5 | *Knowing the exact train the person was travelling on* | 6.4% | mean excluding NR | 6.2% |
| 6 | *Accuracy of other information about the journey undertaken* | 5.3% | mean excluding NR | 5.2% |
| 7 | *Incidence rate* | 4.5% | median | 4.4% |
| 8 | *Response rate* | 5.7% | mean | 5.6% |
| 9 | *Speed of generating topline results* | 5.1% | mean | 5.0% |
| 10 | *Weighting efficiency* | 5.9% | mean | 5.8% |
| 11 | *Cost* | 6.2% | mean excluding RDG | 6.0% |
| 12 | *Cutting data by different time periods* | 6.6% | median | 6.4% |
| 13 | *Interview length* | 4.8% | median | 4.7% |
| 14 | *Ability to recontact participants* | 2.9% | median | 2.8% |
| 15 | *The ability to merge with other data* | 6.4% | median | 6.2% |
| 16 | *Practicability/Feasibility* | 7.0% | Mean + 0.4% | 6.8% |

The rebased weightings (adding to 100%) were sorted into descending order and appear to fit into five groups as outlined in table 9 below.

**Table 9: Scoring criteria in descending order based on rebased initial recommendations weight**

| # | Criteria | Initial recommendation (rebased to 100%) | Gap | SOR |
|---|---|---|---|---|
| 1 | *Coverage of the required universe* | 9.0% | 0 | y |
| 2 | *Ability to generate a random representative sample* | 8.9% | 0.1% | y |
| 4 | *Ability to generate required sample sizes of key subgroups* | 8.7% | 0.2% | y |
| 3 | *Ability to generate large samples (scalability)* | 8.4% | 0.3% | y |
| 16 | *Practicability/Feasibility* | 6.8% | 1.6% | |
| 12 | *Cutting data by different time periods* | 6.4% | 0.4% | |
| 5 | *Knowing the exact train the person was travelling on* | 6.2% | 0.2% | y |
| 15 | *The ability to merge with other data* | 6.2% | 0.0% | y |
| 11 | *Cost* | 6.0% | 0.2% | |
| 10 | *Weighting efficiency* | 5.8% | 0.3% | y |
| 8 | *Response rate* | 5.6% | 0.2% | |
| 6 | *Accuracy of other information about the journey undertaken* | 5.2% | 0.4% | |
| 9 | *Speed of generating topline results* | 5.0% | 0.2% | y |
| 13 | *Interview length* | 4.7% | 0.3% | |
| 7 | *Incidence rate* | 4.4% | 0.3% | |
| 14 | *Ability to recontact participants* | 2.8% | 1.6% | |

The delineation between the groups could be debated but it is clear that the four criteria at the top have somewhat higher weightings than the rest and the one at the bottom is much lower than any other. The rest in the middle are a bit more bunched but a new group was created whenever there is a drop of at least 0.4% from the previous criterion. The top four are all in Appendix B of the Statement of Requirements (SOR) as are half the third group and just one of the bottom two groups. Criterion #10 in the second group was not in the SOR but was added as any methodology being considered needs to be feasible.

The conclusion is that the overall weightings that have been applied are consistent with the criteria defined in the SOR which gives more credence to using this weighting process to combine the scores for each methodology and generate an overall score for each methodology as a basis for ranking them and selecting the top scores.

## 3.5 Final weights provided

The Technical Working Group discussed these weights at their meeting on 13th October 2022 and requested a change to the weight given for "weighting efficiency". The rationale for this change was as follows:

**"Our rationale for giving this a higher weight is because the weighting efficiency will impact the statistical reliability of findings. We think this is particularly important for differences required for significance when comparing findings over time at aggregate and key sub-group levels (e.g. TOC) and comparing findings between key subgroups (e.g. TOCs). This links to the survey requirements to produce data which can be used to evaluate the impact of investments, improvements and service changes and monitor/manage performance of operators and others responsible for rail customer experience, as well as ensuring all stakeholders, including the public, can have confidence in the findings. The technical group agreed that the weighting efficiency criterion should be commensurate with criterion 2 (ability to generate a random representative sample) and criterion 4 (ability to generate required sample sizes of key sub groups) because it will in part determine these and the confidence intervals of survey results. So this means it should be increased to fall into the blue section please (so at or above 8.4)."**

Increasing one weight reduces all the others and to accommodate weighting efficiency with a weight equal to the lowest in the original top (blue) category results in the following table of final weights (table 10).

**Table 10: Scoring criteria and final weights**

| # | Criteria | Final weights | Gap | SOR |
|---|---|---|---|---|
| 1 | *Coverage of the required universe* | **8.7%** | **0.0%** | y |
| 2 | *Ability to generate a random representative sample* | **8.6%** | **0.1%** | y |
| 4 | *Ability to generate required sample sizes of key subgroups* | **8.5%** | **0.2%** | y |
| 3 | *Ability to generate large samples (scalability)* | **8.2%** | **0.3%** | y |
| 10 | *Weighting efficiency* | **8.2%** | **0.0%** | y |
| 16 | *Practicability/Feasibility* | **6.6%** | **1.5%** | |
| 12 | *Cutting data by different time periods* | **6.3%** | **0.4%** | |
| 5 | *Knowing the exact train the person was travelling on* | **6.1%** | **0.2%** | **y** |
| 15 | *The ability to merge with other data* | **6.1%** | **0.0%** | **y** |
| 11 | *Cost* | **5.9%** | **0.2%** | |
| 8 | *Response rate* | **5.4%** | **0.5%** | |
| 6 | *Accuracy of other information about the journey undertaken* | **5.0%** | **0.4%** | |
| 9 | *Speed of generating topline results* | **4.8%** | **0.2%** | **y** |
| 13 | *Interview length* | **4.6%** | **0.3%** | |
| 7 | *Incidence rate* | **4.3%** | **0.3%** | |
| 14 | *Ability to recontact participants* | **2.8%** | **1.5%** | |

These final weights have been used in the process to combine the scores for the 16 criteria and generate an overall weighted score for each of the potential methodologies.

# 4 Outline of phase 3 process

This stage of the process involves listing all methodologies which might possibly be used to measure rail customer satisfaction. The next (fourth) phase involves scoring each of the methodologies against the agreed list of criteria. Variations in a methodology might affect how it is scored e.g., one method might be face to face intercepts allowing passengers to opt for either paper or online completion at their choice. A second method might be to prioritise online completion and only offer paper as a last resort. The second method would be quicker and cheaper than the first but might generate a different response rate. As such, when the two methods are scored against the criteria, they would be likely to receive different scores for some of the criteria.

The list of methodologies therefore needs to include all major variations for any mixed methodology i.e., one where more than one method of data collection is being used. The resulting list is therefore quite long.

An initial list of methodologies was created using the following input:

- Methods used in the past by the GB rail industry including those used in NRPS and Wavelength.
- Methods used in ad hoc projects including the Future of Rail project and the Mixed Methodology trial conducted on behalf of Transport Focus.
- Methods used by TOCs.
- Methods used in the rail industry in other countries.
- Methods used for other transport modes.
- Methods not used to our knowledge.

Documentation for each methodology used was sourced to enable information on how the methodology performed on each criterion to be extracted (this will be used in phase 4).

## 4.1 First iteration

A first iteration was sent to the client on 27th September 2022. The list was developed using the approach described above and comprised 34 possible methodologies outlined in table 11 below.

**Table 11: Initial list of possible methodologies 27th September 2022**

| # | Source | Data collection | Journey assessed | Used in ... |
|---|---|---|---|---|
| 1 | Face to face intercepts at station | Paper only | Current | NRPS 1999-2003, Nantes, SNCF (TER) |
| 2 | Face to face intercepts at station | Paper and email natural split | Current | Early TF Multimethod survey |
| 3 | Face to face intercepts at station | Email with paper backup | Current | Tested in TF Multimethod survey |
| 4 | Face to face intercepts at station | Email with telephone backup | Current | Tested in TF Multimethod survey |
| 5 | Face to face intercepts at airport | Tablet | Current | French airports |
| 6 | Face to face intercepts on board | Paper only postal return | Current | Early BPS, Nantes, French regional rail |
| 7 | Face to face intercepts on board | Paper only can collect on board plus postal | Current | Mid BPS, NRPS boosts, DfT strike survey |
| 8 | Face to face intercepts on board | Paper postal only return and email | Current | Not used to our knowledge |
| 9 | Face to face intercepts on board | Paper postal and collect on board and email | Current | Later BPS, NRPS boosts |
| 10 | Face to face intercepts at | Paper only | Current | NRPS 2004-2010 |

| # | Source | Data collection | Journey assessed | Used in ... |
|---|---|---|---|---|
| | station and on board | | | |
| 11 | Face to face intercepts at station and on board | Paper and email at customer choice | Current | NRPS 2011-2020 |
| 12 | Face to face intercepts at station and on board | Paper and email prioritising email | Current | Not used to our knowledge |
| 13 | Face to face intercepts at station + social media | Paper for intercepts, posts on social media on station satisfaction (Talkwalker) | Current and last journey in past 7 days | SNCF - Gares & Connexions |
| 14 | Online panel | Online survey | Last journey in past 7 days | Wavelength, TF weekly survey |
| 15 | Online panel plus social media recruitment | Online survey | Last journey in past 7 days | Future of Rail (IRPS) |
| 16 | Online panel + face to face | Online survey with paper backup | Last journey in past 7 days | Not used to our knowledge |
| 17 | Telephone | Telephone | Last journey in past 7 days | Not used to our knowledge |
| 18 | Online panel + telephone | Online survey with telephone backup | Last journey in past 7 days | Tested in TF Multimethod survey |
| 19 | Online panel | Online survey | Journey today/yesterday | Not used to our knowledge |
| 20 | Online panel plus social media recruitment | Online survey | Journey today/yesterday | Not used to our knowledge |
| 21 | Online panel + face to face | Online survey with paper backup | Journey today/yesterday | Not used to our knowledge |
| 22 | Telephone | Telephone | Journey today/yesterday | Not used to our knowledge |

| # | Source | Data collection | Journey assessed | Used in ... |
|---|---|---|---|---|
| 23 | Online panel + telephone | Online survey with telephone backup | Journey today/yesterday | Not used to our knowledge |
| 24 | Customer database | Email invitation to online survey | Journey today/yesterday | SNCF - TGV satisfaction survey |
| 25 | Customer database | Email invitation to online survey | Last journey in past 7 days | Disabled Italian passengers, some TOCs |
| 26 | Customer database plus social media recruitment | Email invitation to online survey | Last journey in past 7 days | Toulouse, Eurobahn |
| 27 | Postal database e.g., PAF | Postal | Last journey in past 7 days | Used by TF for their Logistics and Coach survey |
| 28 | App/website | Pop up with survey link | Current | Not used to our knowledge |
| 29 | App/website | Pop up with survey link | Last journey in past 7 days | Some TOCs + PIDD (passenger disruption) |
| 30 | App/website | Pop up with survey link | Journey today/yesterday | Not used to our knowledge |
| 31 | Posters in stations and on trains | Open link to online survey | Current | Station: HS1 - St Pancras sat survey |
| 32 | Train operator's social media channels + online panel | Open link to online survey | Last journey in past 7 days | PIDD (passenger disruption) |
| 33 | Train operator's social media channels + online panel | Open link to online survey | Current and journey today/yesterday | Not used to our knowledge |
| 34 | Third party reservation system | Ad in reservation confirmation with link to online survey | Last journey in past 7 days | Not used to our knowledge |

Feedback was provided on this first iteration that the range of methodologies considered should include:

- Those where intercepts could be supplemented with online, Computer Assisted Telephone Interviewing (CATI) etc.
- Methodologies should be considered that are paper with email backup.
- should consider the use of a QR code, and with paper options consider what is given to respondents such as a paper version of the questionnaire with a link online, a flyer etc. and how visible paper copies of the questionnaire are.
- When considering combined face to face and panel surveys what format the face to face element would take.
- Although it is unlikely to be possible for the sake of inclusion methodologies using customer database should be included as well as hybrid approaches and checking that mixed methodologies could result from the process.

## 4.2 Second iteration

A second iteration was provided to the weekly group participants on 7th October 2022 taking into account the various points raised and following a lot of internal debate (table 12). Some methodologies are new ones in relation to the original list whilst some show specific amendments made to points raised. The methodologies have been re-ordered to provide a more logical sequence.

**Table 12: Second Iterations of proposed methodologies 7th October 2022**

| # | Source | Data collection | Journey assessed | Used in ... |
|---|---|---|---|---|
| 1 | Face to face intercepts at station | Paper postal only | Current | NRPS 1999-2003, Nantes, SNCF (TER) |
| 2 | Face to face intercepts at station | Paper postal and email natural split | Current | Early TF Multimethod survey |
| 2a | Face to face intercepts at station | Paper postal, email and QR code natural split | Current | Early TF Multimethod survey |
| 3 | Face to face intercepts at station | Email with paper postal backup | Current | Tested in TF Multimethod survey |
| 3a | Face to face intercepts at station | QR code with paper postal backup | Current | Not used to our knowledge |
| 3b | Face to face intercepts at station | Email and QR code with paper postal backup | Current | Not used to our knowledge |
| 3c | Mixed source - Face to face intercepts at station + online panel | Email and QR code with paper postal backup plus online panel | Journey today/yesterday | Suggested by NR |
| 3d | Mixed source - Face to face intercepts at station + online panel | Email and QR code with paper postal backup plus customer database | Journey today/yesterday | Not used to our knowledge |
| 3e | Mixed source - Face to face intercepts at station + online panel | Email and QR code with paper postal backup plus App/website | Journey today/yesterday | Not used to our knowledge |
| 3f | Mixed source - Face to face intercepts at | Email and QR code with paper postal backup plus social media recruitment | Journey today/yesterday | Not used to our knowledge |

| # | Source | Data collection | Journey assessed | Used in ... |
|---|---|---|---|---|
| | station + online panel | | | |
| 3g | Mixed source - Face to face intercepts at station + online panel | Email and QR code with paper postal backup plus telephone | Journey today/yesterday | Not used to our knowledge |
| 4 | Face to face intercepts at station | Email with telephone backup | Current | Tested in TF Multimethod survey |
| 5 | Face to face intercepts at station | Face to Face tablet | Current | French airports |
| 6 | Face to face intercepts at station | Email and QR code natural split | Current | Not used to our knowledge |
| 8 | Face to face intercepts on board | Paper postal only | Current | Early BPS, Nantes, French regional rail |
| 8a | Face to face intercepts on board | Paper postal and email also offered natural split | Current | Not used to our knowledge |
| 8b | Face to face intercepts on board | Paper postal and QR also offered natural split | Current | Not used to our knowledge |
| 8c | Face to face intercepts on board | Paper postal and email and QR code also offered | Current | Not used to our knowledge |
| 8d | Face to face intercepts on board | Paper postal and email is prioritised | Current | Not used to our knowledge |
| 8e | Face to face intercepts on board | Paper postal and QR code is prioritised | Current | Not used to our knowledge |
| 8f | Face to face intercepts on board | Paper postal and email and QR code are prioritised | Current | Not used to our knowledge |
| 9 | Face to face intercepts on board | Paper postal and collect on board | Current | Mid BPS, NRPS boosts, DfT recent survey |
| 9a | Face to face intercepts on board | Paper postal and collect on board and email natural split | Current | Later BPS, NRPS boosts |
| 9b | Face to face intercepts on board | Paper postal and collect on board and QR code natural split | Current | Not used to our knowledge |

| # | Source | Data collection | Journey assessed | Used in ... |
|---|---|---|---|---|
| 9c | Face to face intercepts on board | Paper postal and collect on board and QR code and email natural split | Current | Not used to our knowledge |
| 9d | Face to face intercepts on board | Paper postal and collect on board and email prioritised | Current | Not used to our knowledge |
| 9e | Face to face intercepts on board | Paper postal and collect on board and QR code prioritised | Current | Not used to our knowledge |
| 9f | Face to face intercepts on board | Paper postal and collect on board and QR code and email prioritised | Current | Not used to our knowledge |
| 9g | Face to face intercepts on board | Email and QR code natural split | Current | Not used to our knowledge |
| 10 | Face to face intercepts at station and on board | Paper postal only | Current | NRPS 2004-2010, NRTS (but during COVID) |
| 11 | Face to face intercepts at station and on board | Paper postal and email natural split | Current | NRPS 2011-2020 |
| 11a | Face to face intercepts at station and on board | Paper postal and QR code natural split | Current | Not used to our knowledge |
| 11b | Face to face intercepts at station and on board | Paper postal and email and QR code natural split | Current | Not used to our knowledge |
| 12 | Face to face intercepts at station and on board | Paper postal and email prioritising email | Current | Not used to our knowledge |
| 12a | Face to face intercepts at station and on board | Paper postal and QR code prioritising QR code | Current | Not used to our knowledge |
| 12b | Face to face intercepts at station and on board | Paper postal and email and QR code prioritising digital methods | Current | Not used to our knowledge |
| 12c | Face to face intercepts at station and on board | Email and QR code natural split | Current | Not used to our knowledge |

| # | Source | Data collection | Journey assessed | Used in ... |
|---|---|---|---|---|
| 13 | Mixed source - Face to face intercepts at station + social media | Paper for intercepts, analysis of social media posts on code natural station satisfaction (Facebook, Instagram...) | Current and last journey in past 7 days | SNCF - Gares & Connexions |
| 14 | Online panel | Online survey | Last journey in past 7 days | Wavelength, TF weekly survey |
| 16 | Mixed source - Online panel + face to face intercepts at station | Online survey with paper backup | Last journey in past 7 days | Not used to our knowledge |
| 17 | Telephone | Telephone | Last journey in past 7 days | Not used to our knowledge |
| 18 | Mixed source - Online panel + telephone | Online survey with telephone backup | Last journey in past 7 days | Tested in TF Multimethod survey |
| 19 | Online panel | Online survey | Journey today/yesterday | Not used to our knowledge |
| 20 | Mixed source - Online panel + social media recruitment | Online survey | Journey today/yesterday | Not used to our knowledge |
| 22 | Telephone | Telephone | Journey today/yesterday | Not used to our knowledge |
| 23a | Mixed source - Online panel + telephone | Online survey with telephone backup | Journey today/yesterday | Not used to our knowledge |
| 23b | Mixed source - Online panel + social media recruitment | Online survey | Journey today/yesterday | Not used to our knowledge |
| 24 | Customer database | Email invitation to online survey | Journey today/yesterday | SNCF - TGV satisfaction survey |
| 25 | Customer database | Email invitation to online survey | Last journey in past 7 days | Disabled Italian passengers, some TOCs |
| 26 | Mixed source - Customer database plus social media recruitment | Email invitation to online survey | Last journey in past 7 days | Toulouse, Eurobahn |
| 27 | Postal database e.g., PAF | Postal | Last journey in past 7 days | Used by TF for their Logistics and Coach survey |
| 28 | App/website | Pop up with survey link | Current | Not used to our knowledge |

| # | Source | Data collection | Journey assessed | Used in ... |
|---|---|---|---|---|
| 29 | App/website | Pop up with survey link | Last journey in past 7 days | Some TOCs + PIDD (passenger disruption) |
| 30 | App/website | Pop up with survey link | Journey today/yesterday | Not used to our knowledge |
| 31 | Posters in stations and on trains | Open link to online survey | Current | Station: HS1 - St Pancras sat survey |
| 32 | Mixed source - Train operator's social media channels + online panel | Open link to online survey | Last journey in past 7 days | PIDD (passenger disruption) |
| 33 | Mixed source - Train operator's social media channels + online panel | Open link to online survey | Current and journey today/yesterday | Not used to our knowledge |
| 34 | Third party reservation system | Ad in reservation confirmation with link to online survey | Last journey in past 7 days | Not used to our knowledge |
| 35 | Social media | Online survey (free found sample on social media) | | |

This second iteration was signed off by the weekly meeting on 10th October 2022 and has fed into the proforma for phase 4 – "Scoring of all these possible methodologies against all the criteria agreed in phase 1 and the production of an aggregate score for each methodology using the weightings derived in phase 2."

# 5 Outline of phase 4 process

This stage of the process involves five experts scoring all the potential methodologies derived in phase 3. For each methodology the process is as follows:

- Provide a score from 0-100 for each of the criteria derived in phase 1.
- Aggregate the scores using the weights agreed in phase 2, to generate an overall weighted score for that methodology.

The 59 possible methodologies identified in phase 3 involve 16 different solus methodologies either on their own or in combination with others. An initial assessment of the 16 solus methodologies was made to help the scoring process. Each individual then provided a score for each of the methodologies for one of the criteria and then repeated this for the other criteria. Each individual rated the criteria in a different order to minimise any order effects. However, the methodologies were always presented in the same order as this enabled similar methodologies to be grouped (e.g., all the methodologies involving face to face intercepts at stations were in adjacent rows of the pro forma).

Each individual's scores on a criterion were ranked, from 1 being the method with the highest score to 59 being the method with the lowest. Ranking minimises the impact of the range used by each individual and allows a better comparison of results across the individuals undertaking the scoring.

The methodologies were sorted on their aggregate score for each individual. Analysis of the results has been undertaken to identify outliers and the individual providing the outlier score asked to comment on why their score appears to be so different from others. Arithmetic means, medians, minima, maxima and spread were calculated for each of the criteria across all the methods.

## 5.1 Descriptions of the solus methodologies

A pro forma was created to enable each of the 16 solus methodologies to be described and a semantic scale created for each. As an example, the information for the response rate criterion is shown below in table 13 (similar information was provided for each of the sixteen criteria). The information on response rate in the right column is a mixture of actual response rate and rating on the semantic scale.

**Table 13: Solus Methodologies with data collection method and possible response rate**

| # | Solus methodology | Data collection method | Semantic scale - Response rate |
|---|---|---|---|
| | | | 100 Very high |
| | | | 75 Fairly high |
| | | | 50 Moderate |
| | | | 25 Fairly low |
| | | | 0 Very low |
| | | | What response rate is the method likely to generate |
| 1 | F2F intercepts at station - Paper | Paper | 20% (NRPS, Multimethod) |
| 2 | F2F intercepts at station - Email | Online | 25% (Multimethod) |
| 3 | F2F intercepts at station - QR code | Online | 25% (Multimethod) |
| 4 | F2F intercepts on board - Paper | Paper | 25% (NRPS) |
| 5 | F2F intercepts on board - Email | Online | 30% (extrapolated from at station NRPS) |
| 6 | F2F intercepts on board - QR code | Online | 30% (estimated based on NRPS and Multimethod |
| 7 | Online panel | Online | Low response rate - Wavelength is about 7% |
| 8 | Telephone | Telephone | 10% (typical telephone response rate) |
| 9 | Customer database | Online | Moderate/low response rate - Typical customer database yields 10% or so |
| 10 | Postal database e.g., PAF | Online / Paper | 3% is typical postal response rate for postal survey |
| 11 | App/website | Online | Moderate/low response rate |
| 12 | Posters in stations and on trains with QR code | Online | Low response rate |
| 13 | Train operator's social media channels | Online | Low response rate |
| 14 | Third party reservation system | Online | Moderate/low response rate - Typical customer database yields 10% or so |

| # | Solus methodology | Data collection method | Semantic scale - Response rate |
|---|---|---|---|
| 15 | Social media recruitment | Online | Low response rate based on social media user universe but higher based on those clicking on the ad |
| 16 | F2F intercepts at station (disembarking) | Tablet | Moderate response rate |

The contents of this proforma were discussed at a meeting on 14th October 2022 where the agency team, peer reviewer and client were present.

## 5.2 Process for individuals to follow

The criteria were scored in a different order for each individual completing the task. This approach minimises the impact of any order effects. The order used was as follows:

- The first assessor scores the criteria in the order on the proforma.
- The second assessor scores in reverse order.
- The third assessor scores in order but starts in the middle.
- The fourth assessor scores in reverse order but starts in the middle.
- The fifth assessor scores in a random order.

This process ensures that each criterion appears early and late in the process a similar number of times and appears before its adjacent criterion roughly the same number of times as it appears after. Using these variations of order in which the criteria have been scored reduces any order effects which might arise due to tiredness, boredom or increasing speed and possibly reducing quality during the scoring process. The individuals undertaking the scoring were advised to take breaks to further reduce these impacts but to nevertheless ensure each criterion was scored fully during one period.

Individuals were encouraged to look at the solus pro forma for the criterion they were rating beforehand and to have it easily to hand to refer to as they completed their scores for the 59 possible methods, either by printing out the relevant columns to have in front of them or by opening the solus methodology proforma in another window as they scored all the methodologies.

The evaluation pro forma had an instructions tab as follows:

**Figure 1: Evaluation pro forma instructions**

| |
|---|
| The proforma in the second tab needs to be completed for each of the 16 criteria listed in Cols E-T of the proforma tab |
| The process for each of the 16 criteria is as follows: |
| |
| Familiarise yourself with the comments given for the criterion in the relevant column of the "solus method summary" tab |
| If you are rating the criterion in col E, this information will be in column E of the solus criteria tab |
| Rate each of the criteria in the order with which you have been assigned |
| When rating a mixed methodology, look at the comments given for each constituent part of that methodology to help generate your score |
| Complete all the scores in a maximum of 10 minutes |
| When you have finished, sort the criteria into descending order of scores using all of rows 8-67 with the header in row 8 and the column being sorted row as the sort key |
| Satisfy yourself that that sorted list of methods looks sensible and take a maximum of 5 minutes to do this |
| Sort back into the original order using all of rows 8-67 and the key in col A as the sort variable |
| Save the spreadsheet and repeat for the other criteria |
| |
| It is perfectly acceptable and indeed desirable to take breaks between scoring criteria. However, try to score each criterion in a single pass - don't take breaks going down a column unless essential |
| |
| When you have completed all the criteria, save the file and email to [redacted] by 12:30 pm on Monday 17th October at the latest |

## 5.3 Initial analysis – outliers and consistency

As expected there were some variations in the scores given by the experts. To understand the reasons for this, each person was provided with their data and the overall average and asked to explain what might have led to their divergence. The methods and criteria which were outliers for each person are shown in Annex A together with their responses. The feedback provides useful context to the scoring process for each person but we have not used it to amend any of the scores.

Before generating results from the data, we ran a number of tests to assess whether the five individuals had generated similar scoring profiles. Firstly, we looked at the spread of scores across the 59 methodologies given for each criterion by each of the assessors, outlined in table 14.

**Table 14: Expert assessors' scores**

*(minus (-) sign indicates lowest score given, plus sign (+) indicates highest score given)*

| Journey assessed | Expert 1 | Expert 2 | Expert 3 | Expert 4 | Expert 5 |
|---|---|---|---|---|---|
| Coverage of the required universe | 85 (-) | 90 | 95 | 100 (+) | 97 |
| Ability to generate a random representative sample | 70 | 60 (-) | 90 | 80 | 93 (+) |
| Ability to generate large samples (scalability) | 75 | 80 | 50 (-) | 90 | 97 (+) |
| Ability to generate required sample sizes of key subgroups | 62 (-) | 90 (+) | 72 | 87 | 90 (+) |
| Knowing the exact train the person was travelling on | 62 | 80 (+) | 15 | 10 (-) | 29 |
| Accuracy of other information about the journey | 60 | 70 (+) | 10 (-) | 23 | 40 |
| Incidence rate | 88 | 90 | 100 (+) | 90 | 87 (-) |
| Response rate | 62 (-) | 85 | 94 (+) | 69 | 89 |
| Speed of generating topline results | 73 (-) | 100 (+) | 100 (+) | 100 (+) | 93 |
| Weighting efficiency | 40 (-) | 40 (-) | 90 (+) | 40(-) | 45 |
| Cost | 90 (-) | 92 | 95 | 100 (+) | 94 |
| Cutting data by different time periods | 20 (+) | 0 (-) | 0 (-) | 0 (-) | 0 (-) |
| Interview length | 40 (-) | 75 | 80 (+) | 75 | 70 |
| Ability to recontact participants | 75 (-) | 100 (+) | 95 | 75 (-) | 90 |
| The ability to merge with other data | 5 | 20 | 25 | 3 (-) | 95 (+) |
| Practicability/Feasibility | 30 | 75 (+) | 55 | 75 (+) | 18 (-) |

For some criteria the spread of scores across the 59 methodologies was very consistent:

- coverage of the required universe,

- ability to generate required sample sizes of key subgroups,
- incidence rate,
- response rate,
- speed of generating topline results,
- cost,
- cutting data by different time periods,
- interview length,
- ability to recontact participants.

For some other criteria, there was just one outlier e.g., Expert 3 for 'ability to generate large samples (scalability)' and 'weighting efficiency' and Expert5 for the 'ability to merge with other data' and 'practicability/feasibility'.

For other criteria three experts had similar scores and the two other experts had similar scores but different to the former three.

- Ability to generate a random representative sample.
- Knowing the exact train the person was travelling on.
- Accuracy of other information about the journey.

For these criteria there was a suggestion that individuals were rating slightly different things from their colleagues.

In general, we do not see this as a major issue – people will have different views about things – and averaging results and subjecting them to the challenges at the workshop enabled a wide range of views to be taken into account in creating the scores for each of the criteria and the overall weighted scores.

## 5.4 Initial analysis – results

Now we had scores for each of the criteria for each method, we could aggregate them into a single weighted score using the profiles derived in phase 2. The methods could be ranked and comparisons made both overall and for each of the assessors. If there was consistency in the top ranked methods between the assessors, we could be reasonably confident that these methods were indeed rated more highly.

The results were presented at the workshop on 24th October. Some of the results presented at the workshop generated debate.

There was some discussion of the scores with regard to which methodologies would generate random samples. There was some doubt that online panels would generate larger samples and their scalability. Questions were asked about the weighting will be used and the importance of the field trials to work out the approach. Questions were asked about how practicality was being considered, was it what was possible or what was logistically easy. How would a QR code work in practice and what effect did different methodologies have on older passengers' response rate. Other logistical issues were raised such as how quickly could data be received and what methodologies would speed that up. Additionally, there was some discussion about how additional value could be achieved with regard to cost. The group asked to consider the minimum and maximum samples and also what the journey of reference could be for respondents under different methodologies i.e. the journey that day of the respondent or a journey in the last seven days.

The main conclusions were as follows:

- The top 31 methods are all face to face intercepts of one sort or another.
- There are several variants of mixed methods – face to face intercepts at station + digital methods coming next in order of the overall aggregate scores.
- The first online panel methodology is in 38th place and uses the evaluation of journey made today/yesterday (rather than in the last 7 days).
- There are some methods which, at present, could be disregarded as principal methods due to their low scores although they might be more relevant once digital services increase coverage (apps, geo location etc.) and these are shown on the next page.

There was a view discussed and broadly agreed at the workshop that there was little point in putting very similar methodologies into the field trials. Rather, it would be a better approach to select different methodologies with their preferred data collection approaches

and have some of the resource in each field trial devoted to testing subtle nuances of the main method. For example, intercepting passengers on board a train and collecting data via a natural split of email and QR code emerged as the top methodology.

However, some concern was expressed that not having a paper option at all might exclude older passengers. The weighted approach puts email/QR code ahead of email/QR code/paper as the inclusion of paper increases cost and involves a lot longer for topline results to be produced and in the weighted approach these drawbacks more than outweigh the benefit of a slightly more representative sample. The direct comparison can be seen in table 15.

**Table 15: Comparison of weighted score for email/QR code data collection and email/QR/Paper based data collection methodologies**

*(minus (-) sign indicates the score was the lowest given, plus sign (+) indicates the score was the highest given)*

| **Data collection**<br><br>**Journey assessed** | Face to face intercepts on board<br><br>Email and QR code natural split<br><br>Current | Face to face intercepts on board<br><br>Paper postal and collect on board and QR code and email prioritised<br><br>Current |
|---|---|---|
| Weighted total | **83.56** | **79.34** |
| Rank | **1** | **5** |
| Coverage of the required universe | **97.8** | **97.8** |
| Ability to generate a random representative sample | **69.0 (-)** | **70.6 (+)** |
| Ability to generate large samples (scalability) | **98.8** | **98.8** |
| Ability to generate required sample sizes of key subgroups | **92.4 (-)** | **93.6 (+)** |
| Knowing the exact train the person was travelling on | **99.8** | **99.8** |
| Accuracy of other information about the journey undertaken | **93.6 (+)** | **92.0 (-)** |
| Incidence rate | **99.8** | **99.8** |
| Response rate | **80.8 (-)** | **81.4 (+)** |
| Speed of generating topline results | **94.0 (+)** | **29.4 (-)** |
| Weighting efficiency | **67.6 (-)** | **68.4 (+)** |
| Cost | **53.0 (+)** | **41.0 (-)** |
| Cutting data by different time periods | **92.0** | **92.0** |
| Interview length | **81.0 (+)** | **72.2 (-)** |
| Ability to recontact participants | **55.0 (-)** | **57.2 (+)** |
| The ability to merge with other data | **81.0 (+)** | **80.4 (-)** |
| Practicability/Feasibility | **70.2 (+)** | **66.2 (-)** |

As a result of the initial analysis, the following considerations were agreed:

- The three face to face intercept methods should feature in the field trials.
- Variations in the methods used to collect the data have an impact and email/QR code using the natural split scores highest.

- These could be the core methodologies with some shifts allowing paper or short link completion to confirm whether adding paper or a short link provides any advantages.
- The best method which majors on a digital approach is an online panel using journey today/yesterday which is in 38th place and we recommend this features in the trials.
- Telephone, postal, app based, customer database based, and social media can be ignored as principal methods at this point in time due to their poor scores, but some may be worthy of consideration in the future or as subsidiary methods in a mixed approach.

Following discussion, the combined station/on board method was replaced with a method using databases, as it was felt this might be more of a real possibility in future, with digital ticket sales becoming the norm. The impact of a combined at station/on board methodology could always be assessed by combining results from the two individual approaches.

The four final methods were thus agreed at the workshop as follows:

**Method 1**

- face to face intercepts on board,
- email/QR code natural split, and
- Some of the fieldwork to incorporate a paper backup.

**Method 2**

- face to face intercepts at station,
- email/QR code natural split, and
- some of the fieldwork to incorporate a paper backup.

**Method 3**

- Commercial online panel,
- Online survey, and
- Last journey in past seven days but allowing analysis of journey today/yesterday.

**Method 4**

- Third party or TOC customer database,
- Online survey, and
- Last journey in past seven days but allowing analysis of journey today/yesterday.

In summary, the four methods initially agreed for the field trials are outlined below in table 16.

**Table 16: four methods agreed for field trials**

| Method | Respondent source | Data collection | Options | Journey assessed |
|---|---|---|---|---|
| 1 | Intercepts on board | Email/QR code | Some paper | Current journey |
| 2 | Intercepts at station | Email/QR code | Some paper | Current journey |
| 3 | Commercial panel | Online | None | Last 7 days |
| 4 | Customer database | Online | None | Last 7 days |

## 5.5 Methodology recommendations for field trials

Following the workshop on 24th October 2022 and the ensuing discussion, an outline note was put together summarising the suggested approach for the field trials.

**Field trials:**

At the end of the session, we agreed on investigating three further distinct methodologies (customer database was taken out of scope initially):

1. Face to face intercepts on board with email and QR natural split.
2. Face to face intercepts at station with email and QR natural split.
3. Commercial online panel.

The scores for those methods agreed through phases 1-4 of the method review are outlined in table 17 below.

**Table 17: Criteria scores for each agreed method**

| Source | **Face to face intercepts at station** | **Face to face intercepts on board** | **Online panel** |
|---|---|---|---|
| Data collection | **Email and QR code natural split** | **Email and QR code natural split** | **Online survey** |
| Journey assessed | **Current** | **Current** | **Last journey in past 7 days** |
| Weighted score | 81 | 84 | 65 |
| Weighted rank | 3 | 1 | 45 |
| Coverage of the required universe | 89 | 98 | 65 |
| Ability to generate a random representative sample | 67 | 69 | 81 |
| Ability to generate large samples (scalability) | 99 | 99 | 48 |
| Ability to generate required sample sizes of key subgroups | 92 | 92 | 47 |
| Knowing the exact train the person was travelling on | 91 | 100 | 62 |
| Accuracy of other information about the journey undertaken | 93 | 94 | 57 |
| Incidence rate | 88 | 100 | 29 |
| Response rate | 72 | 81 | 47 |
| Speed of generating topline results | 94 | 94 | 97 |
| Weighting efficiency | 70 | 68 | 34 |
| Cost | 53 | 53 | 79 |
| Cutting data by different time periods | 92 | 92 | 92 |
| Interview length | 81 | 81 | 68 |
| Ability to recontact participants | 55 | 55 | 73 |
| The ability to merge with other data | 81 | 81 | 81 |
| Practicability/Feasibility | 68 | 70 | 93 |

## 5.6 Expected monthly sample size per month

1. Face to face intercepts with email and QR on board

We expect 100 to 180 respondents per TOC per rail period. We recommend booking 10 shifts per TOC per rail period which means we expect 10-18 completed questionnaires per three hour shift.

These figures are just indicative and will depend on the time of the day and day of the week the shifts are conducted. For the field trials we recommend trying to achieve 100 completes per TOC – as this would give us enough data to draw robust conclusions. For the field trials we would recommend adding on boosts if needed once we have seen the returns after the first few shifts.

2. Face to face intercepts with email and QR at station

In a face to face intercept at the station scenario, it is much harder to estimate the number of recruits/completes per shift. This depends a lot on the sample plan and which stations are to be included in the field trials.

Larger stations will result in more recruits and more completes.

There are also some regional differences with some areas where completion rates are generally lower (for example Scotland).

There will be a natural fallout from the sample. If this does not meet targets, boost shifts would need to be undertaken at Network Rail (NR) stations (for any sub analysis specific to NR stations). In the field trials we can also look at adding in data from face to face shifts which intercept passengers on board trains that depart from NR stations, as this would probably be more efficient than additional at station shifts.

3. Online commercial panel

In table 18 below is the expected sample size per TOC per rail period if we reach 5,000 respondents. They were calculated based on previous studies including Wavelength and the IRPS. There will be no quotas, so the number of respondents per TOC will naturally fallout.

**Table 18: expected sample size per TOC per rail period based on assumed sample of 5,000 respondents**

| Last 7 days | Today/yesterday |
|---|---|
| Avanti West Coast: 130-230<br>c2c: 70-140<br>Chiltern Railways: 70-140<br>Cross Country: 90-190<br>East Midlands Railway: 130-230<br>Elizabeth line: 90-290<br>Gatwick Express: 20-120<br>Grand Central: 20-120<br>Great Northern: 90-190<br>Great Western Railway: 260-360<br>Greater Anglia: 160-260<br>Heathrow Express: 10-60<br>Hull Trains: 20-80<br>London North Eastern Railway LNER: 110-210<br>London Northwestern Railway: 50-140<br>London Overground: 220-340<br>Lumo: 5-50<br>Merseyrail: 50-130<br>Northern: 380-480<br>ScotRail : 110-250<br>South Western Railway: 350-550<br>Southeastern: 440-550<br>Southern: 310-410<br>Thameslink: 180-320<br>Transpennine Express: 100-200<br>Transport for Wales: 70-160<br>West Midlands Railway: 90-160 | Avanti West Coast: 40-71<br>c2c: 26-53<br>Chiltern Railways: 24-48<br>Cross Country: 29-62<br>East Midlands Railway: 59-72<br>Elizabeth line: 31-103<br>Gatwick Express: 8-45<br>Grand Central: 6-36<br>Great Northern: 34-71<br>Great Western Railway: 83-115<br>Greater Anglia: 57-93<br>Heathrow Express: 5-30<br>Hull Trains: 7-27<br>London North Eastern Railway LNER: 37-70<br>London Northwestern Railway: 16-44<br>London Overground: 92-142<br>Lumo: 3-25<br>Merseyrail: 15-38<br>Northern: 131-166<br>ScotRail : 39-89<br>South Western Railway: 114-180<br>Southeastern: 145-181<br>Southern: 109-144<br>Thameslink: 62-110<br>Transpennine Express: 33-65<br>Transport for Wales: 21-49<br>West Midlands Railway: 30-53 |

Our recommendation for the field trial is to open up to travel in the last seven days and to ask when their most recent trip was. We will then be able to analyse and compare the profile and answers of those who travelled today/yesterday versus those who travelled in the last seven days.

# 6 Sampling and weighting for the field trials

A number of data sources are available for use in sampling and weighting the data for rail customer experience surveys. This section describes the various sources providing pros and cons for each and also comes up with recommended options for sampling going forward and for weighting the field trials data. The experience gained from weighting the field trials data, which is likely to be an iterative process, will then help define the final weighting regime for a future continuous survey.

The sources evaluated are as follows:

- Office for Rail and Road (ORR) data on the number of passengers at each station.
- ORR data on the number of passengers for each TOC.
- The RDG Electronic Rail Timetable containing details of every scheduled train service.
- The MOIRA database which contains estimated passenger numbers of each service.
- The LENNON ((Latest earnings Networked Nationally Overnight) database which is driven by the national ticket sales system.

Each is described below. Following the description of each method, a summary of the pros and cons of each is provided.

The section of the report then recommends solutions for at station and on board sampling moving forward together with a recommended weighting regime for the field trials data.

## 6.1 ORR data on passenger numbers at each station

Data is produced annually by Steer for the ORR. This dataset is published around November each year for the year ending the previous March. So, the latest data for the year to end March 2022 was published in November 2022 and the latest and data (ORR, *Table 1410:Passenger entries, exits and interchanges by station (annual), Great Britain, April 2021 to March 2022*, (Nov 2022)) is shown here. A detailed report on how the data is generated is produced regularly (ORR, *Estimates and Station Usage: Quality and Methodology report*, (2023)) and is shown here. The key information about this data source consists of estimates of the total number of people:

- Travelling from or to the station (entries and exits); and
- Changing trains at the station (interchanges).

The estimates of entries and exits are further split by ticket type (full price, reduced price and season tickets). There is also a range of station attribute information included, e.g., geographic data. Time series of entries and exits and interchanges by station are available from April 1997.

Estimates of station usage are primarily based on sales data from LENNON, the rail industry's ticketing and revenue system. This is supplemented with some local ticketing data. Below is a list of all the data sources used to create the estimates of station usage:

Entries and exits:

- LENNON, Transport for London (TfL) data and train operator data (Gatwick Express and Stansted Express) as an input to the MOIRA2.2 base matrix,
- local ticketing data from Passenger Transport Executives (PTEs),
- manual station counts, and
- Heathrow Express ticketing data.

Interchanges:

- Central Allocations File (CAF).

Pros of this data source

1. It is comprehensive covering nearly all stations in GB (just two small stations were omitted from the most recent release).

2. It incorporates information from a wide variety of sources of ticket sales.

3. It has been an Official Statistic recognised as such by the ONS since 2020.

4. It has trend data measured on a consistent basis.

Cons of this data source

1. It is out of date by the time it is published so does not reflect any short/medium term changes in travel behaviour.

2. It does not split passenger numbers by TOC at station level (although in conjunction with the RDG Electronic Rail Timetable this is possible).

3. It does not include data for recently opened stations nor for new operators calling at existing stations (for example there is currently limited data on passenger numbers for the Elizabeth line).

## 6.2 ORR data on passenger numbers for each TOC

The ORR also generates estimated annual passenger journeys by operator (TOC) (ORR, *Table 1223: Passenger journeys by operator, Great Britain, April 2011 to December 2023* (March 2024)) and the latest data is shown here. This data is published quarterly about three months after the end of the quarter. Annual passenger journeys are also split by sector and ticket type but not within operator. The data allows the profile of passenger journeys by TOC to be calculated.

Pros of this data source

1. It is comprehensive covering nearly all stations in GB (just two small stations were omitted from the most recent release).
2. It incorporates information from a wide variety of sources of ticket sales.
3. It has been an Official Statistic recognised as such by the ONS since 2020.
4. It has trend data measured on a consistent basis.
5. It is reasonably up to date.

Cons of this data source

1. It does not split passenger numbers by TOC at station level (although in conjunction with the RDG Electronic Rail Timetable this is possible).

## 6.3 The RDG Electronic Rail Timetable

The electronic timetable data can be downloaded from Data Download | data.atoc.org using the timetable feed. To access this data, the user needs to register to both the National Rail Data Portal at National Rail Data Portal and then through the LINKS option on the data download menu to the RSP feeds. Registration is free to individuals or companies with a legitimate need to use the data.

Downloading the timetable feed generates a large zip file containing eight text data files. The filename reflects the date of the download (files downloaded on 22/12/22 have a file name ttisf585 whilst those downloaded on 28/12/22 have a filename ttisf592). The structure of the eight files is well described in this file:

RSPS5046_timetable_information_data_feed_interface_specification.pdf (raildeliverygroup.com)

The eight text files cover the following content:

**Table 19: Eight text data files from RDT electronic timetable**

| Filename | Content |
|---|---|
| ttisf585.ZTR | Z Trains file |
| ttisf585.REJ | TTIS Rejects |
| ttisf585.SET | Common Interface File Set |
| ttisf585.FLF | Fixed Links |
| **ttisf585.MCA** | **Basic Timetable Detail** |
| **ttisf585.MSN** | **Master Station Name File** |
| ttisf585.ALF | Additional Fixed Links |
| ttisf585.TSI | TOC Specific Interchange Times |

The two emboldened files are those used – basic timetable data and station names. The link between the two is a TIPLOC code, which identifies the station name in the master station file and is used to identify locations in the Basic Timetable file.

From these files it is possible to create a number of derived files:

- A file with data for each train service including departure time, origin station, destination station, days run, start date, end date and operator.
- A file with a record for each station where each service calls including station name, days run and operator.

The first file can be used to select a systematic sample of services from all those that are scheduled to run across a given time period (this was done for the pilot study). This does not take into account passenger volumes of each train and so will select low volume services at the same intensity as high volume services. The latter file can include origin station, destination station and every intermediate station where the service calls. By aggregating all services for a station, it is possible to calculate the number of services each week which call at a station both in total and split by operator. The profile of all services by operator can be calculated and compared to the profile of passengers from the ORR data at operator level. The comparison gives an average load factor for each operator which is shown at Annex C.

For each station on the ORR database, the split of services by operator from the timetable can be applied, together with the average load factor, to generate the split of passengers by operator for each station. This was used in the pilot study to select the at station sample.

Pros of this data source

1. It is comprehensive covering all train services scheduled to operate from now onwards.
2. It incorporates information from all train operators.
3. It is completely up to date.

Cons of this data source

1. It does not have any data on likely passenger volumes for each service.
2. It is a very large file (5 million+ records) requiring bespoke analytic tools.
3. It uses a different station identifier (TIPLOC codes) from other data sources (which tend to use Three Letter Codes).

## 6.4 The MOIRA database

MOIRA is a system which generates passenger estimates and revenue for each train service. the Department for transport provided a file (based upon December 2022 data) which shows passenger numbers boarding and alighting for each station for each train service on the database. This database can be aggregated in several ways:

- For each station, to show the number of passengers boarding for each TOC.
- For each train service, to show the total number of passengers on the train.

There are three files, one for an average weekday, one for Saturdays and one for Sundays. The weekday file does not take into account the number of days on which a service runs, but this is likely to be a small issue given that most will be five days per week. Annual estimates from MOIRA have been computed from adding five times the weekday passenger numbers to the Saturday and Sunday numbers and then multiplying this total by 50 to turn a week into an estimated annual figure.

Using this approach, the annual passenger estimate from MOIRA was reasonably consistent with figures in the ORR data.

There are some stations on the MOIRA database which are not on the ORR database. Most of these are new stations, but there are some which have different TLC codes and some which are dealt with differently on the two systems (e.g., for Heathrow Airport).

The data relating to passenger numbers for each TOC at each station has been compared to that produced using the ORR data and number of services calling at each station derived from the electronic rail timetable. There are differences in the profile by TOC for each station, which reflect the older nature of the ORR data (year to March 2022 compared to December 2022 for MOIRA) and the differing loadings of passengers on each TOC. 0).

Looking at data at a train service level, this is complicated by the fact that MOIRA does not contain the unique train identifier which is on the electronic rail timetable but rather has a train code made up from the departure time, origin TLC and destination TLC. This train code can be generated from the electronic rail timetable data to see how well the services on MOIRA match those on the timetable. Matching services that were scheduled in April/May 2023 and used for the sample for the pilot with MOIRA data from the December 2022 extract shows a 91% match rate.

Given that MOIRA is not updated that frequently, the above shows that most services on the electronic timetable do have a match to MOIRA and the estimated volumes from MOIRA could be patched into the sample section process so that services could be selected with

probability proportional to the number of passengers. Services that did not match would need some default setting perhaps based upon the TOC average and the length of journey.

The factors used to calculate the number of passengers boarding and alighting at each stop are based on data from before the COVID pandemic. As such, they will probably overstate passenger numbers at weekday peak times and understate passenger numbers at other times.

Pros of this data source

1. It covers most of the train services scheduled to operate in the electronic timetable.
2. It is a source that the industry uses for other activities.
3. It is reasonably up to date (more so than the ORR data) in terms of train services.

Cons of this data source

1. It does not have any data for some train services (around 10%).
2. It is not completely up to date and in particular the factors used to generate passenger number estimates are from pre the pandemic.

## 6.5 The LENNON database

LENNON contains data generated from ticket sales data. It is unlikely to be useful for journey sampling purposes but may provide data that is more up to date than other sources for weighting purposes, as it is updated every day. To assess LENNON, we were provided with an extract for two TOCs (Chiltern and LNER) and then a ticket type analysis for each TOC for the latest financial year – April 2022 to March 2023 – to enable LENNON data to be compared with that produced by Steer for the ORR. The Steer data used for the ORR data includes ticket sales from sources outside the national ticketing system and the comparison allows us to see which TOCs are most affected by these exclusions.

Of the 24 TOCs examined, seven had a different number of passenger journeys in ORR data compared to LENNON. For these seven TOCS the passenger journeys in the ORR data ranged from -1% to +37% of the passenger journeys indicated in LENNON data. The TOCS that had different figures are below.

- East Midlands Railway,
- Govia Thameslink Railway,
- Greater Anglia,
- London Overground,
- Merseyrail,
- West Midlands Trains, and
- Heathrow Express (totally excluded from national ticketing data).

Apart from the first and last TOCs, the others are all in PTE areas or the TfL area and as the comments in Annex E show are the TOCs most likely to be affected by ticket sales exclusions. We know that ticket sales outside the national ticketing system are significant for some TOCs and the above analysis confirms this.

The difference between the ORR data and the LENNON data can be used to generate factors which represent total journeys divided by journeys measured on the LENNON database. In principle this factor could be applied to updated LENNON journey estimates to cater for the ticket sales not included. We recommend consideration of this process to update estimates of total TOC journeys from LENNON analysis for relevant fieldwork periods.

We were provided with LENNON analysis for the period of the field trials and the factors were applied to the relevant TOCS.

Once the factors were applied the comparison between the estimated profile using LENNON data and that from ORR shows rises for Avanti West Coast, CrossCountry, Merseyrail, Northern and Transport for Wales and falls for GTR and London Overground.

For ticket type, the comparison between ORR and LENNON for the period April 2022 to March 2023 is as follows:

**Table 20: Comparison of ORR and LENNON ticket type profiles April 2022 to March 2023**

| | profile ORR | Profile LENNON |
|---|---|---|
| Franchised ordinary ticket Advance | 6.77% | 6.31% |
| Franchised ordinary ticket Anytime or Peak | 31.94% | 32.23% |
| Franchised ordinary ticket Off Peak | 46.11% | 46.65% |
| Franchised ordinary ticket Other | 0.55% | 0.40% |
| Franchised Season ticket | 14.62% | 14.41% |
| | | |
| | 100.00% | 100.00% |

As with the operator profile, the LENNON data for the year April 2022 to March 2023 excludes ticket sales from local sources and the resulting profile is very similar to that shown on the ORR website. As a result, we believe it would be permissible to use updated LENNON ticket type data as a means to assess if the journey profiles from the Rail Experience Survey needed any weighting by this factor.

LENNON actually supplies a more granular breakdown of ticket type than that used to compare with ORR data as follows:

APEX SINGLE/RETURN
BRITRAIL/CONTINENTAL
CHEAP DAY SINGLE
FIRST ADVANCE PURCHASE
FIRST CHEAP DAY RTN/DAY TRVLCARD
FIRST REDUCED
FIRST RETURN
FIRST SEASONS 180-359 DAYS VB2B
FIRST SEASONS 91-180 DAYS VB2A
FIRST SEASONS ANNUAL
FIRST SEASONS UP TO 90 DAYS VB1
FIRST SEASONS WEEKLY
FIRST SINGLE
INCLUSIVE TOURS
MISCELLANEOUS
NON PASSENGER/RAIL TRAVEL
NON SPECIFIC SPG
OTHER ADVANCE PURCHASE SGL/RTN
OTHER REDUCED SINGLE/RETURN
REFUNDS BY FLOW ORDINARY
REFUNDS BY FLOW SEASONS

ROVER TICKETS
SAVER
SLEEPER SUPPLEMENT
STANDARD FLEXI SEASON
STANDARD RETURN
STANDARD SINGLE
STD CHEAP DAY RTN/DAY TRVLCARD
STD SEASONS 180-359 DAYS VB2B
STD SEASONS 91-180 DAYS VB2A
STD SEASONS ANNUAL
STD SEASONS UP TO 90 DAYS VB1
STD SEASONS WEEKLY
SUPER ADVANCE SINGLE/RETURN
SUPERSAVER

LENNON data from these categories, or combinations of them to align with the Rail Experience questionnaire, could thus be used to assess whether any additional weighting by ticket type is required.

## 6.6 Summary of pros and cons of the various sources

Table 21 below indicates how well each source performs against a range of criteria.

**Table 21: Summary of sample data performance against different criteria**

| | ORR station data | ORR TOC data | RDG Electronic timetable | MOIRA | LENNON |
|---|---|---|---|---|---|
| **Comprehensive** | **Y** | **Y** | **Y** | **Y** | **Y** |
| **Industry acceptance** | **Y** | **Y** | **y** | **Y** | **Y** |
| Official statistic | Y | Y | N | N | N |
| Trend data | Y | Y | N | N | Y |
| **Up to date** | **N** | **N** | **Y** | **Y** | **Y** |
| Station data | Y | N | Y | Y | N |
| TOC data | N | Y | Y | Y | Y |
| New stations/lines | N | N | Y | Y | Y |
| Passenger numbers | Y | Y | N | Y | Y |
| TLC Codes | Y | Y | N | N | N |

No source is best on all criteria, but we have emboldened those that seem the most critical to us: comprehensive, industry acceptance, and up to date.

On this basis, the latter three sources seem preferable to use of the ORR data, where the main concern is data being out of date.

## 6.7 Conclusion - at station sampling

The MOIRA data is more up to date than the ORR data and enables passenger numbers for each TOC at each station to be produced from a current industry data source (whereas the current method uses a bespoke procedure that assumes passenger services per train are constant for a given TOC). It seems that using MOIRA data would be a distinct improvement and allow more up to date and more robust data to be used to select the sample. It needs to be borne in mind that estimated passenger numbers for weekday peak times are probably overstated and those at other times understated due to the lack of updating of some of the load factors applied within MOIRA.

This data source could also be used for weighting as it is possible to construct day of week profiles and time of day profiles for each TOC and also use some station or station size weighting.

All this data can be analysed by TOC and used for weighting purposes. This seems a distinct improvement on the profiles that TOCs used to provide for NRPS with no apparent provenance.

## 6.8 Conclusion - on board sampling

Adding MOIRA data onto the estimated number of passengers for each service to the RDG electronic rail timetable and using sampling proportionate to the estimated number of passengers is preferable to the current method of sampling services at random. The number of low volume services that are selected would be reduced. Estimates would need to be made for those services on the electronic timetable but not on MOIRA and these could be based upon the TOC and the journey length.

This data source could also be used for weighting as it is possible to construct day of week profiles and time of day profiles for each TOC and also use some line of route weighting such as TOC building blocks.

## 6.9 Weighting the main field trials data

If we are using MOIRA data for sampling, it would be consistent to use it for weighting. MOIRA can be used to provide targets for each TOC by weekday/weekend and time of day. Early analysis of the field trials data at Annex D shows that the key satisfaction measures do vary by day of week and time of day. We have therefore constructed the following dayparts so that weighting by these dayparts does counter any response rate differences.

- weekday - morning peak (trains starting between 06:00 and 08:59).
- weekday - evening peak (trains starting between 16:00 and 18:59).
- weekday – late (trains starting from 19:00 onwards).
- weekday – other (off-peak – starting before 06:00 and between 09:00 and 15:59).
- Saturday.
- Sunday.

Weighting for each TOC could be done for groups of stations separately (a building block approach), to ensure that any bias towards larger stations is corrected. This would mirror the approach used in NRPS. We have therefore divided the stations for each TOC into roughly four equal bands after sorting by number of passengers (so the top one or two stations are in band 1, the next largest in band 2 and a larger group of smaller stations in band 3 and especially small ones in band 4).

We recommend using the profiles that emerge from MOIRA analysis covering these day parts and station size bands to weight the field trials data.

Up to date estimates of the passenger numbers for each TOC can be provided by LENNON and applying a factor to take account of ticket sales not covered by the national ticketing system.

Weighting will be applied separately for the at station data and the onboard data and ideally for each by the base approach and the option tested (QR code at station and paper back up on train). One of the purposes of testing the two intercept approaches – at station and on train – is to compare the results and weighting each to the same profile will help in that task. The at station approach generates respondents who did not use one of the 12 selected TOCs and we recommend excluding these respondents from the main analysis. We will then be comparing passengers using the same 12 TOCs.

When combining the 12 TOCs together, we recommend using the passenger profile by TOC which comes from the adjusted LENNON data for the fieldwork period.

Early data for the field trials (comparing questionnaires complete to footfall data) suggests that there is no great response bias by age but there is by gender. We should therefore weight data demographically by age group and gender from the footfall counts. This weighting should be undertaken at the total sample level as the numbers for an individual TOC are likely to be too small (and for the at station approach it is not possible to separate out the footfall counts for specific TOCs at some stations). A decision will then need to be made on the frequency and volume of footfall counts to ensure this process is valid.

Initially we will not weight by ticket type or journey purpose and hope that the random sampling approach we have used to select stations and trains for sampling and weighting by daypart, station size band and demographics will yield the correct ticket type profile. We will compare the profile with that from the LENNON analysis and if the ticket type profile generated by this weighting regime is significantly different, we should also weight by ticket type at TOC level.

We need to remain aware that LENNON data does not include all ticket sales and this is particularly a problem for certain TOCs (London Overground, Merseyrail, Heathrow Express and indeed any other TOC where substantial numbers of tickets are sold outside the channels that LENNON incorporates).

It should be borne in mind that any extra variable used in the weighting process reduces the effective sample size, sometimes dramatically, if the sample profile is significantly different from the universe profile. This is particularly the case when looking at national data, as TOCs are sampled disproportionately to ensure each TOC has a robust sample size.

## 6.10 Weighting the interventions

We do not recommend weighting the data for the various interventions, as the sample sizes will be quite small. For the interventions the analysis will compare the end profile of the survey versus our ideal target and seeing how closely or not we land next to our representative weighting criterion. So, we would not weight but use the weights to compare the profiles and see if there is one intervention generating concerning respondent profiles.

## 6.11 Weighting the data for managed stations

Network Rail has a specific requirement to produce robust data for each of the stations it manages. The sample sizes for each station from the national sample may not be large enough to ensure this for each station and in NRPS this was resolved by undertaking boost shifts at stations where the sample size was insufficient. In addition, the profile of passengers at each station by daypart and by TOC may not match that expected. It is important to note that an NR managed stations report does not have to use exactly the same approach as for the overall TOC analysis.

For the field trials analysis, we recommend including all respondents in the managed stations reports, not just those for the selected TOCs (of course once all TOCs are included in the pilot survey, this will not be an issue). For the pilot survey, to meet Network Rail objectives, it may be necessary to boost the main sample to both achieve a required minimum sample size for each station and to generate the required passenger profile by daypart and by TOC.

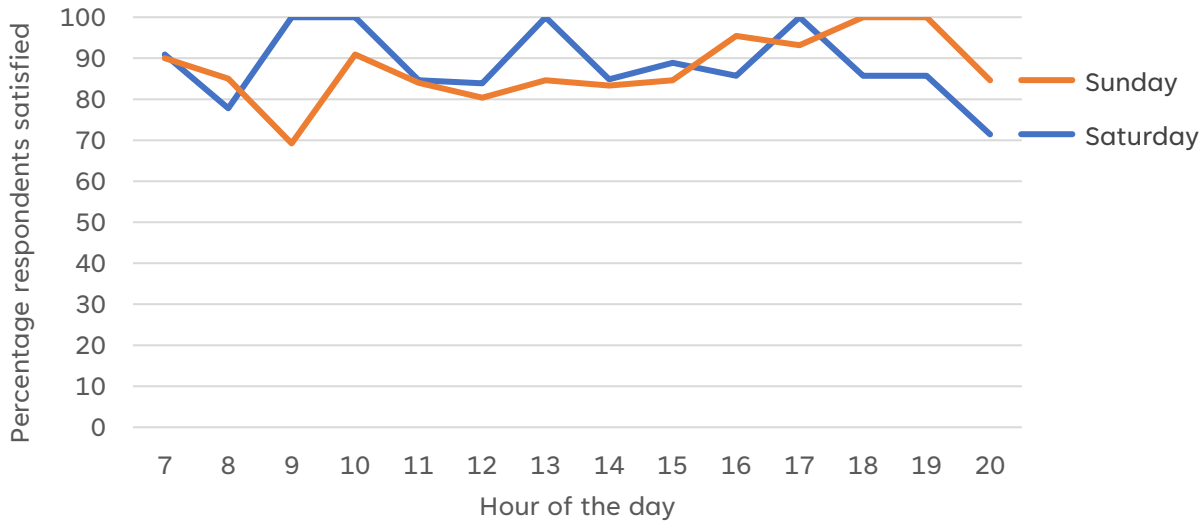
## 6.12 Specific questions addressed

*Is there an argument for different day parts at the weekend, particularly Saturday, where the daytime market will be very different from the evening market? Similarly for Friday, does this need specifically separating out from other weekdays?*

To use different dayparts, there needs to be evidence that key metrics vary more by these different dayparts than those originally suggested. We also need to have universe data to weight to any new subdivisions.

Satisfaction on Saturdays and Sundays does not vary in any systemic way by time of day. There is a peak at 11.00-12.00 but this is based upon quite small samples (53 at 11.00 and 60 at 12.00) so probably not significantly different from surrounding times (sampling error on a percentage of 90% will be around +/-10% on these sample sizes). The adjacent hours have lower satisfaction and it is hard to discern any clear pattern across the day. There is thus no strong argument for subdividing these days.

**Figure 2: Customer overall journey satisfaction by time of day on Saturdays and Sundays**



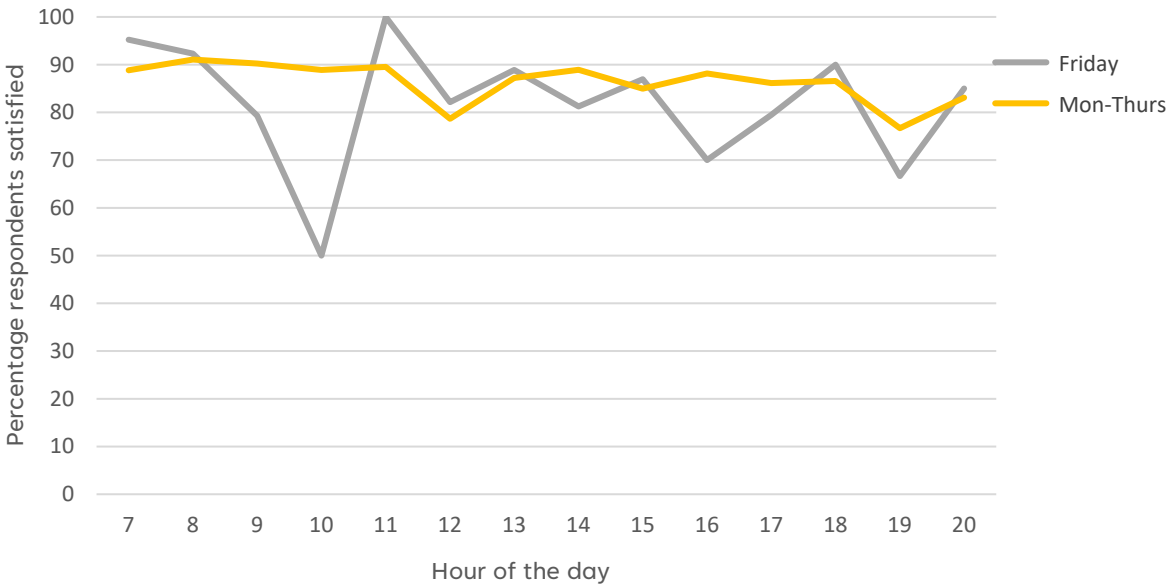

Source: customer experience survey field trials data

The pattern of overall satisfaction is similar for Fridays as compared to other weekdays. There are patterns that satisfaction in the evening peak tends to be lower than at other times, confirming that peak time should be used as a separate daypart, as recommended.

There is also no universe data for passenger numbers by hour of day for Fridays. MOIRA only has data on a typical weekday, so even if satisfaction on Fridays showed a different pattern from other weekdays, there would be no universe data on which to weight this.

The effect of any different performance on individual days can be managed by sampling similar numbers of passengers on each weekday and within each day controlling the sample by time of day.

**Figure 3: Comparison of customer satisfaction by time of day on Fridays and weekdays**




Source: customer experience survey field trials data

*The point around Network Rail sample – could this be elaborated and some rough numbers put against it, as it plays directly into the sample discussion?*

This will depend upon the methodology adopted. In the field trials, around a quarter of the responses had the journey origin as one of Network Rail's 20 managed stations. This number was affected by the restriction of the sample to 12 TOCs- so for example there were no journeys that originate at London Cannot Street. Once the methodology has been agreed and the initial sample size agreed and selected, we will be able to estimate how many responses will come from each of the 20 Network Rail stations and determine how many top up shifts are likely to be required.

*Has any analysis of the data been done to see if there is any variation we need to control beyond these usual suspects?*

The data tables will analyse key metrics by a number of criteria and this will help identify any other factors worthy of consideration. As highlighted above, any variable used for weighting should satisfy two criteria:

- Key metrics vary significantly by the subgroups of the variable.

- Universe data exists for each subgroup to enable weighting.

It should also be borne in mind that the addition of any further weighting variables will reduce the effective sample size.

# Annex A – Feedback on rationale for scoring outliers

Expert4 - feedback

**Method 13 - Intercepts at station – F2F tablet data collection in situ**

***Coverage***

I believe that coverage would be good because almost all people leaving the station would have recently been on a train journey. I don't foresee much differences in this incidence vs. passengers arriving.

***Speed of generating results***

As the data are collected via the tablet at the station, the results are instantly available for analysis if they can be uploaded via a mobile network. Otherwise at the end of the day via Wifi which I consider still fast.

***Cost***

You have highlighted cost but my ranking is only very slightly different to everyone else 52.5 vs 54 (1.5 points). I don't consider this difference consequential.

***Interview length***

Everyone else put this as the longest method. I believe that in person will be faster or very similar to administer vs. telephone and I believe that it will be faster to complete than by paper.

***Ability to recontact participants***

I believe that the rapport that the interviewer builds up during the interview is more likely to lead to the respondent providing either their email address or a telephone number than most other methods where we don't automatically collect contact information.

**Method 23 -Intercepts on board – Paper postal, paper collect on board, email**

My weighted average is only 1.1 points outside the others' average and the ranking is only five away.

***Accuracy of the information***

I think with some of the data being collected by interviewers this should lead to more accurate information being provided and less misunderstanding of questions or mistakes filling in the survey.

***Speed of generating results***

4.5 points difference is pretty much on a par with others and there are very many options with a similar mix of survey completions. So, a small change in one rating could have a big impact on rankings. I suspect this is the case for this one.

***Weighting efficiency***

I have provided very similar ratings for all of the in-person recruitment approaches. So, I think I have a lot of ratings that are similar with a small change resulting in a large change in rank.

***Cost***

I don't believe administering the questionnaire on board will be less efficient vs not and therefore the same number of shifts would be required. I have put this in as mid ranking and I believe this is the case.

***Ability to merge with other data***

I think I rated pretty much all the methods the same on this one, hence why it is mid tier (as they all probably are). I think the difference in this case is likely to be a very small variance in rating leading to have very large difference in ranking.

I think we need to consider a blend of both absolute rating differences and also ranking differences as in a few cases highlighted above we are amplifying discrepancies which are in fact reasonably well aligned.

Expert3 feedback

**Method 18 – Face to face intercept on board – Paper postal and email and QR code also offered**

***Ability to generate a random representative sample***

I rated this method highly on this criterion because for me this method is one of the most inclusive and will not exclude any key subgroups. It allows respondent to pick the method of completion they feel most comfortable with. There is no completion method bias and this will lead to a good random representative sample. The Transport Focus Multimethod project

has shown that allowing people to choose their method of completion delivers a very representative sample.

***Speed of generating topline results***

This is far from being a quick method to generate topline results compared to other methods as paper is included. I have given this method a zero on this criterion. It is the same rating I gave for all other methods where paper was involved. My ranking is biased by the fact that 35 methodologies are getting the same result (zero).

***Ability to recontact participants***

It is not the most straight forward methodology to achieve this criterion. As it is a mix mode some elements are good for this criterion (email – as you capture the email address at the recruitment stage – the only thing you need is permission to recontact) and others not so good (paper and QR). This method is better than if email is not part of the method. With Paper and QR, on top of asking for permission your need to ensure the email address capture is error free. For QR you can ensure that the online script validates the format and the consistency (if you ask respondents to type their email address twice). With paper you have no control over the email address field. A mix mode like this one is therefore mixed on this criterion; better than just Paper or Paper and QR but worse than Email, or Email and QR. Here again, I rated all the Paper, QR and Email methodologies the same on this criterion (x 9 methodologies)

***The ability to merge with other data***

For this criterion this method is not particularly great. It scored 70. 50 other methods scored the same. It does not offer any additional benefits for this criterion.

***Practicability/Feasibility***

This is not a high ranking methodology for this criterion. I rated this 70 on this criterion. This is the same score for 23 other methodologies. The main difficulties here are around:

- Sorting out the permission for the fieldwork (same to all other face to face intercept)
- Sampling, briefing and monitoring fieldwork (same to all other face to face intercept)
- Printing and processing the paper questionnaires (same to all other paper postal methods)
- Setting up the email invitation and reminders (same to all other email methods)

- The QR code here does not had any added complexity.

**Method 19 – Face to face intercept on board – Paper postal and email is prioritised**

For this method it looks like I rated this lower than the others – below is why:

***Accuracy of other information about the journey undertaken***

This is not a very discriminating criterion. I only gave four different scores, all rating high (100 – 90). This method rated 95 – mainly because of the paper element – there is no control over which questions the respondent will answer and we might receive some incomplete questionnaires.

***Interview length***

Because of the paper element, you are limited in the number of questions you can ask, and you cannot really have different blocks of questions rotating. One option would be to only do this for email respondents as it is something easy to do for the online script, but it would not be a clean solution as all of those answering the paper questionnaire will not be asked the same questions. I rated all the methodologies with a paper element the same regardless of which other methods they are mixed with.

***Ability to recontact participants***

It is not the most straight forward methodology to achieve this criterion. As it is a mix mode email is good for this criterion (as you capture the email address at the recruitment stage – the only thing you need is permission to recontact) and Paper not so good. This method is better than if QR was part of the method. For QR you need ensure that the online script validates the format and the consistency (if you ask the respondent to type his email address twice). Here the only difficulty is with paper for which you have no control over the email address field. A mix mode like this method is therefore mixed on this criterion; better than just Paper or Paper and QR but worse than Email, or Email and QR.

***Practicability/Feasibility***

This is not a high ranking methodology for this criteria. I rated this 70 on this criterion. This is the same score for 23 other methodologies (including method 18 above). The main difficulties here are around:

- Sorting out the permission for the fieldwork (same to all other face to face intercept)
- Sampling, briefing and monitoring fieldwork (same to all other face to face intercept)

- Printing and processing the paper questionnaires (same to all other paper postal methods)

- Setting up the email invitation and reminders (same to all other email methods)

**Method 24 – Face to face intercept on board - Paper postal and collect on board and QR code natural split**

***Ability to generate a random representative sample***

I rated this method highly on this criterion because I think it is probably the most inclusive methodology and will not exclude any key subgroups. It allows respondent to pick the method of completion they feel most comfortable with and also removes the issue of having to post the questionnaire back if this was a blocker for the paper route. There is no completion method bias and this will lead to a good random representative sample.

***The ability to merge with other data***

For this criterion this method is not particularly great. It scored 70. 50 other methods scored the same. It does not offer any additional benefits for this criterion.

***Ability to recontact participants***

It is not the most straight forward methodology to achieve this criterion. As it is a mix mode some elements are good for this methodology (email – as you capture the email address at the recruitment stage – the only thing you need is permission to recontact) and others not so good (paper and QR). This method is better than if email is not part of the method. With Paper and QR, on top of asking for permission your need to ensure the email address capture is error free. For QR you can ensure that the online script validates the format and the consistency (if you ask the respondent to type his email address twice. On paper you have no control over the email address field. A mix mode like this method is therefore mixed on this criterion; better than just Paper or Paper and QR but worse than Email, or Email and QR. Here again, I rated all the Paper, QR and Email methodologies the same on this criterion (x 9 methodologies)

On the other criteria with -4 difference – I don't feel the ranking difference is large enough to comment.

**Method 13 – Face to face intercept at station – Face to Face tablet**

***Accuracy of other information about the journey undertaken***

The accuracy will be very good as the interview will take place at the station just after the journey happened. It will be one of the best methods to collect 'hot feedback as it will happen just as they finished their journey, they are rating. Accuracy will be very high. Other methods have a lag time between the journey and the completion.

***Response rate***

This is the method that will deliver the best response rate. Almost all passenger who will start the survey will complete it as this is interviewer led. Any one that stops and agree to take part will then do it. For the other intercept methods, passengers might say they will do it but then will not take part once they have access to the survey. This method is also better than any online alternative because it reduces a lot the dropout rates.

***Speed of generating topline results***

Because the data is inputted directly on the tablet, it is synced automatically (if the tablet is online – which is the recommendation). The data because available as soon as a complete is captured.

***Ability to recontact participants***

This would be fairly easy. As simple as an online survey. The script on the tablet would check for the email format and will validate the consistency (can ask the email twice). The only method that are easier are those for which we already have the contact details (email, customer database...).

***Practicability/Feasibility***

This method is fairly easy to put in place. It is the easiest set up for a face to face intercept method. The only difficulty is to control for interviewer bias. It would be an expensive method but that is not what we are rating here. If the budget is unlimited this is fairly easy to run. One script on the tablet. No need to set invitations, reminders, recruitment questionnaires, printing and scanning...

Expert1 feedback

**Method 12 - Face to face intercepts at station, Email with telephone backup, Current journey**

***Ability to generate a random representative sample***

I rated this method highly on the ability to generate a random representative sample because I thought the email would be able to pick up the younger respondents while the telephone would act as a way to pick up the older respondents. On reflection, being a face to face telephone backup it might not be able to pick up as many older respondents and can understand why this would have been rated lower by others.

***Ability to generate required sample sizes of key subgroups***

I rated highly on sample sizes because I thought that you could boost on the required key groups for both email and telephone and therefore reach the required key groups. I think this comes hand in hand with the coverage of the universe and I felt that both methods together could pick up key subgroups and so adding a boost would help to achieve the sample size needed. As with the first criteria I have commented on for this method, I think as a backup this might not be the case as it would rather need to be a split of the two to achieve a good sample size for each group.

***Response rate***

I rated highly on response rate as I focused on the email element and given this had a higher response rate, I focused on this and inferred that the backup would not impact the responses too much.

***Speed of generating topline results***

For all of these, my rationale for the scoring on generating topline results was that I gave a score of 90 to those that were very fast, and 17 to those where they were very slow (anything with paper). I felt that while the paper took a while to produce topline results, I didn't feel this should warrant a score of zero and rather a midpoint between very and fairly slow.

***Weighting efficiency***

For this, I used the weighting efficiency percentages as a way to score each of the methods. I rated this as 63% as this was a midpoint between 65% for email and 60% for telephone for assessing journeys that are currently taking place and therefore felt this made sense.

**Method 18 - Face to face intercepts on board, Paper postal and email and QR code also offered, current journey**

***Speed of generating topline results***

For all of these, my rationale for the scoring on generating topline results was that I gave a score of 90 to those that were very fast, and 17 to those where they were very slow (anything with paper). I felt that while the paper took a while to produce topline results, I didn't feel this should warrant a score of zero and rather a midpoint between very and fairly slow.

***Ability to recontact participants***

I rated highly on this one because I thought the fact email was offered meant there was a much easier way to recontact participants. Ultimately, this does depend on the uptake on email which I probably overlooked when scoring on recontacts here.

***Practicability/Feasibility***

I rated this method highly on this criterion as I felt that while the postal was a drawback on feasibility that QR and email would be easier enough to offer alongside the paper. I rated all of the methods offered alongside paper postal for face to face intercepts on board the same as felt they would be of equal feasibility due to the same work being conducted for all.

**Method 31 - Face to face intercepts at station and on board, Paper postal and email natural split, Current journey**

***Coverage of the required universe***

I rated both of these two methods the same on their coverage of the require universe as I felt they were similar in what they could provide. I thought this because on board was able to pick up all of the required sample and therefore addressing the objectives, while at the station is slightly less able to. I felt that collectively the two F2F intercepts would be able to pick up a better read of the universe as a whole - more of an addition than an average between the two. I can, however, understand the scores given by others as at the station does have some compromises.

***Knowing the exact train the person was travelling on***

I rated both of the two methods highly on knowing the exact train because I thought that with the sample being taken from on board the train as well that this would pick up the exactness of the train. On second reflection, as with coverage of the universe I would rate this slightly lower now because at the station is limited in that not everyone will be train users and rather visiting the station for other purposes outside of rail travel.

***Speed of generating topline results***

For all of these, my rationale for the scoring on generating topline results was that I gave a score of 90 to those that were very fast, and 17 to those where they were very slow (anything with paper). I felt that while the paper took a while to produce topline results, I didn't feel this should warrant a score of zero and rather a midpoint between very and fairly slow.

***Interview length***

I rated this method highly on interview length because I felt that being able to offer a paper version meant that the length could be quite long and so rated this somewhere halfway between online survey and paper only.

***Ability to recontact participants***

I rated this higher than others as I felt that the email enabled this method to have the ability to recontact more people despite it being a split method. Given it is a natural split, since the initial scoring I may have dropped this down slightly given paper has its disadvantages regarding recontacts.

**Method 32 - Face to face intercepts at station and on board, Paper postal and QR code natural split, Current journey**

***Coverage of the required universe***

I rated both of these two methods the same on their coverage of the require universe as I felt they were similar in what they could provide. I thought this because on board was able to pick up all of the required sample and therefore addressing the objectives, while at the station is slightly less able to. I felt that collectively the two F2F intercepts would be able to pick up a better read of the universe as a whole - more of an addition than an average between the two. I can, however, understand the scores given by others as at the station does have some compromises.

***Knowing the exact train, the person was travelling on***

I rated both of the two methods highly on knowing the exact train because I thought that with the sample being taken from on board the train as well that this would pick up the exactness of the train. On second reflection, as with coverage of the universe I would rate this slightly lower now because at the station is limited in that not everyone will be train users and rather visiting the station for other purposes outside of rail travel.

***Speed of generating topline results***

For all of these, my rationale for the scoring on generating topline results was that I gave a score of 90 to those that were very fast, and 17 to those where they were very slow (anything with paper). I felt that while the paper took a while to produce topline results, I didn't feel this should warrant a score of zero and rather a midpoint between very and fairly slow.

***Ability to recontact participants***

I rated this higher than others as I felt that the email enabled this method to have the ability to recontact more people despite it being a split method. Given it is a natural split, since the initial scoring I may have dropped this down slightly given paper has its disadvantages regarding recontacts.

***Practicability/Feasibility***

I rated this highly as when I was rating I felt that the differences in practicability between emails and QR were not dissimilar. I felt that the same amount of effort and logistics would go into both and so the ratings for both of these mirror each other on this. I still believe this was the right scoring.

Expert2 feedback

**Method 4 - Face to face intercepts at station, Email with paper postal backup, Current journey**

***Ability to generate a random representative sample***

Thought paper backup would balance the bias of email to a small extent.

***Ability to generate required sample sizes of key subgroups***

Both email and paper as part of f2f intercepts were rated as very well on the solus method summary. Hence my rating.

***Speed of generating topline results***

Email is rated as very high speed on the solus method summary sheet and the main approach in this method. Paper is only backup.

***Cost***

Again, email is main approach and rated as moderate on solus summary sheet. Paper which would be the cost driver is only back up.

***The ability to merge with other data***

Both approaches are rated as good on this criterion in solus method summary. Hence my rating. In the context of the way I rated, I feel this is justified.

# Annex B – Methodology for calculating passenger volumes by TOC and station

The following is a description of how ORR and timetable data is used to calculate passenger volumes for each TOC at each station in the national rail network.

**Step 1**

Passenger journey data for each station is taken from the ORR database. This database uses ticket sales data from LENNON supplemented with journey data from a number of other sources that LENNON does not include, principally:

- Data from TfL for London Underground stations that offer national rail services.
- PTE data from sales that are made from sources other than national rail stations.

The data used is half the number of entries and exits plus the number of interchanges. For example, the total annual passenger journeys estimated from London Victoria in the year to April 2022 was 21,684,106 (half the 36,776,338 entries and exits and 3,295,937interchanges).

There are a few new stations that are not on the ORR data, the latest of which relates to the year ending March 2022. For these stations, passenger numbers are estimated by applying the ratio of total national journeys to total services to the number of services run from that station.

**Step 2**

Data from the electronic timetable is used to count how many services each TOC runs from a station in a typical week in the survey period. This is then profiled, so that we estimate what percentage of the services run from a station are by each TOC. At London Victoria, the percentage breakdown of services scheduled to run from the station in March 2022 to April 2023 was as follows: (these percentages are very similar to those generated for NRPS using RailPlanner data in 2016):

| | | |
|---|---|---|
| Southeastern | 32.73% | 32.02% in 2016 |
| Gatwick Express | 6.94% | 10.25% in 2016 |
| Southern | 60.17% | 57.53% in 2016 |
| Thameslink[1] | 0.16 % | 0.19% in 2016 |

Note 1: The occasional Thameslink service calls at London Victoria, hence the small percentage here.

**Step 3**

For each TOC, we know from the ORR data the percentage of all journeys that are on that TOC. From the timetable data, we know what percentage of all services are from that TOC. By comparing the two, we can estimate a journey to passenger converter and apply that to the process. This gives an enhanced breakdown of the estimated passengers for each TOC from London Victoria as follows:

| | |
|---|---|
| Southeastern | 34.16 % |
| Gatwick Express | 6.79% |
| Southern | 58.89 % |
| Thameslink | 0.16 % |

**Step 4**

These profiles are then applied to the total passenger count for the station derived in step 1. Implicitly, the assumption is that the proportion of journeys by TOC from the station is the same as the proportion of number of estimated passengers by TOC from the station. For London Victoria, this results in estimated passenger volumes as follows:

| | |
|---|---|
| Southeastern | 7,407,982 |
| Unmapped (was Gatwick Express) | 1,473,083 |
| Southern | 12,769,220 |
| Govia Thameslink Railway (Thameslink) | 33,821 |

**Step 5**

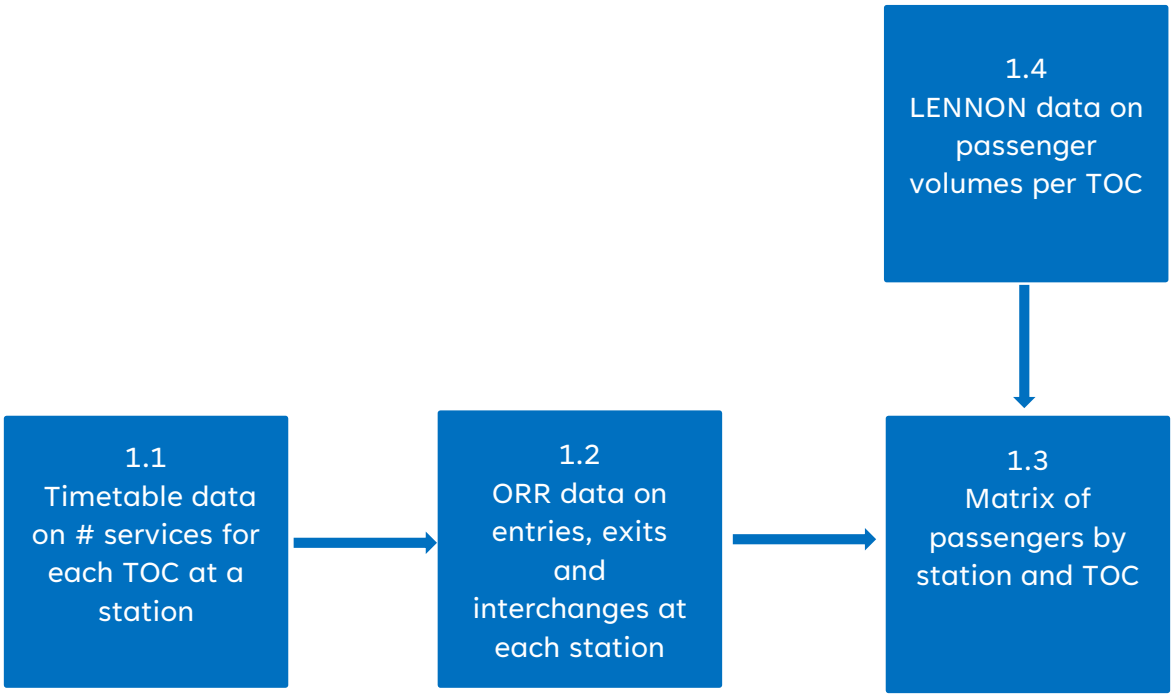
The total estimated passenger journeys for each TOC is computed by adding up the estimate for each station at which the TOC calls. This enables the percentage of journeys on the TOC that start at each station to be calculated.

At the end of this process, we have a set of estimated passenger journeys for each TOC at each station that adds to the station totals and is close to the TOC totals. If each TOC is

sampled independently any variance of the TOC totals is of no consequence as the sampling process is just using data on all the stations for that particular TOC.

The diagram below summarises how this process works.

**Figure 4: NRPS - Stage 1 – Derivation of sampling plan data**



# Annex C – Journey to passenger converter factors

| TOC | Factor |
|---|---|
| Avanti West Coast | 1.74 |
| c2c | 1.77 |
| Caledonian Sleeper | 4.73 |
| Chiltern Railways | 1.30 |
| CrossCountry | 1.69 |
| East Midlands Railway | 1.02 |
| Elizabeth line | 0.85 |
| Gatwick Express | 1.21 |
| Grand Central | 1.46 |
| Great Northern | 1.21 |
| Great Western Railway | 0.95 |
| Greater Anglia | 1.30 |
| Heathrow Express | 1.09 |
| Hull Trains | 1.37 |
| Island Lines | 1.39 |
| London North Eastern Railway | 3.28 |
| London Overground | 1.40 |
| Lumo | 1.61 |
| Merseyrail | 0.50 |
| Northern Trains | 0.55 |
| ScotRail | 0.46 |
| Southeastern | 1.29 |
| Southern | 1.21 |
| SWR | 1.39 |
| Thameslink | 1.21 |
| TransPennine Express | 1.19 |
| Transport for Wales | 0.35 |
| West Midlands Trains | 0.85 |

Typically, TOCs running trains in rural areas including Transport for Wales, Northern and Scotrail have low converters suggesting these services have lower than average passenger numbers.

Commuter TOCs such as c2c, Chiltern, Greater Anglia, London Overground, Southeastern, Southern, SWR and GTR have above average factors, suggesting higher than average passenger numbers. Long distance services including Grand Central, Hull Trains, Avanti West Coast and CrossCountry also have above average converters again suggesting higher than

average passenger numbers. Caledonian Sleeper looks high but it is the ratio of two small numbers.

Most of these results seem logical adding credence to the use of this technique.

# Annex D – Analysis from field trials survey

In this annex, we show selected results from the initial analysis of the field trials survey data, to examine the variation in results by key metrics.

**Table 22: Customer satisfaction variation by time of day**
(Percentage of respondents in stating they were satisfied or very satisfied to the question "Overall, taking everything into consideration, how satisfied or dissatisfied were you with …")

| Hour of departure | Overall satisfaction | The overall punctuality of this service | The frequency of trains running on this route | The overall value for money of the journey you made |
|---|---|---|---|---|
| 0 | 88% | 85% | 81% | 67% |
| 1 | 89% | 81% | 73% | 65% |
| 2 | 85% | 82% | 72% | 58% |
| 3 | 82% | 79% | 71% | 60% |
| 4 | 91% | 83% | 77% | 67% |
| 5 | 88% | 85% | 75% | 69% |
| 6 | 82% | 74% | 68% | 58% |
| 7 | 81% | 86% | 70% | 71% |
| 8 | 88% | 86% | 70% | 56% |
| 9 | 89% | 87% | 74% | 63% |
| 10 | 88% | 85% | 72% | 62% |
| 11 | 88% | 84% | 74% | 63% |
| 12 | 90% | 88% | 79% | 70% |
| 13 | 82% | 85% | 77% | 67% |
| 14 | 88% | 83% | 74% | 68% |
| 15 | 87% | 81% | 76% | 65% |
| 16 | 86% | 78% | 74% | 59% |
| 17 | 86% | 88% | 78% | 66% |
| 18 | 87% | 85% | 79% | 66% |
| 19 | 87% | 76% | 79% | 69% |
| 20 | 81% | 72% | 61% | 61% |
| 21 | 83% | 79% | 64% | 66% |
| 22 | 88% | 85% | 77% | 66% |
| 23 | 88% | 82% | 73% | 62% |
| Total | 87% | 83% | 74% | 63% |

**Table 23: Customer satisfaction variation by day of the week**
(Percentage of respondents stating they were satisfied or very satisfied to the question "Overall, taking everything into consideration, how satisfied or dissatisfied were you with ...")

| Day of week (Monday=1) | Overall satisfaction | The overall punctuality of this service | The frequency of trains running on this route | The overall value for money of the journey you made |
|---|---|---|---|---|
| Mon | 88% | 82% | 74% | 64% |
| Tues | 87% | 83% | 72% | 62% |
| Weds | 86% | 82% | 72% | 63% |
| Thurs | 87% | 83% | 74% | 62% |
| Fri | 85% | 81% | 72% | 59% |
| Sat | 85% | 84% | 79% | 68% |
| Sun | 88% | 85% | 76% | 71% |
| Total | 87% | 83% | 74% | 63% |

**Table 24: Customer satisfaction variation by journey purpose**
(Percentage of respondents stating they were satisfied or very satisfied to the question "Overall, taking everything into consideration, how satisfied or dissatisfied were you with ...")

| Summary journey purpose | Overall satisfaction | The overall punctuality of this service | The frequency of trains running on this route | The overall value for money of the journey you made |
|---|---|---|---|---|
| Unknown | 50% | 50% | 63% | 25% |
| Commuter | 84% | 76% | 65% | 51% |
| Business | 86% | 84% | 76% | 58% |
| Leisure | 89% | 87% | 80% | 73% |
| Other | 85% | 81% | 74% | 69% |
| Total | 87% | 83% | 74% | 63% |

**Table 25: Customer satisfaction variation by gender of respondent**
(Percentage of respondents stating they were satisfied or very satisfied to the question "Overall, taking everything into consideration, how satisfied or dissatisfied were you with ...")

| Gender | Overall satisfaction | The overall punctuality of this service | The frequency of trains running on this route | The overall value for money of the journey you made |
|---|---|---|---|---|
| Male | 84% | 81% | 74% | 61% |
| Female | 89% | 85% | 75% | 66% |
| Another way | 82% | 73% | 62% | 53% |
| Not answered | 72% | 64% | 58% | 40% |
| Total | 87% | 83% | 74% | 63% |

**Table 26: Customer satisfaction variation by age of respondent**
(Percentage of respondents stating they were satisfied or very satisfied to the question "Overall, taking everything into consideration, how satisfied or dissatisfied were you with ...")

| Summary age groups | Overall satisfaction | The overall punctuality of this service | The frequency of trains running on this route | The overall value for money of the journey you made |
|---|---|---|---|---|
| Not answered | 100% | 100% | 67% | 100% |
| 16-34 | 85% | 80% | 71% | 59% |
| 35-54 | 87% | 83% | 73% | 60% |
| 55-64 | 86% | 85% | 75% | 70% |
| 65+ | 90% | 87% | 82% | 81% |
| Other | 81% | 71% | 60% | 57% |
| Total | 87% | 83% | 74% | 63% |

**Table 27: Customer satisfaction variation by daypart of journey**
(Percentage of respondents stating they were satisfied or very satisfied to the question "Overall, taking everything into consideration, how satisfied or dissatisfied were you with ...")

| Daypart | Overall satisfaction | The overall punctuality of this service | The frequency of trains running on this route | The overall value for money of the journey you made |
|---|---|---|---|---|
| Weekday - morning peak | 86% | 83% | 68% | 58% |
| Weekday - evening peak | 87% | 83% | 77% | 63% |
| Weekday - late | 87% | 80% | 69% | 60% |
| Weekday - other | 87% | 83% | 75% | 64% |
| Saturday | 85% | 84% | 79% | 68% |
| Sunday | 88% | 85% | 76% | 71% |
| Total | 87% | 83% | 74% | 63% |

# Annex E – LENNON exclusions

DfT provided information on the types of ticket sales excluded from the LENNON system, as follows:

- Season ticket travel is assumed as you've noted – so whilst the number of journeys on a Weekly Season ticket for example is 10.3 per week, the customer in reality may make less/more journeys than this. When it comes to Annuals and Monthlies – journeys are 'drip fed' into Lennon across the period of ticket duration.

- Nature of allocated travel –in Lennon the route one can take for travel is not always fully known (e.g., Cornwall > Scotland via Any Permitted route – there are a vast number of combinations for travel on this route!), so journeys are 'allocated' to TOCs based on the most likely route a customer will take via ORCATs, but again this is estimated and not reality.

- Open return tickets – whilst the outward date will be stipulated on the ticket, the day the customer chooses to return is unknown so Lennon will just allocate both the outward and return journey to the start date of the ticket. This particularly affects Bank Holiday weekends, where we know customers return on the BH Monday/Tuesday, but the journeys are fully allocated to the outward date (usually Friday of the BH in this case).

- Direction of travel on return/season tickets – there can only be one origin and destination assigned to a ticket in Lennon which will be as per entered in booking stage, so in the case of a return (and also season tickets), all journeys will be assigned to the outward Origin > Destination direction even though the customer will make travel in return direction (for example, a return from Brighton > London Victoria – both journeys will be assigned to Brighton > Victoria, rather than one record for Brighton > Victoria and a separate journey record for Victoria > Brighton.

- PAYG/Contactless journeys are assumed to be based on number of taps) – if this is the case, then this is not aligned to how journeys are created for a normal National Rail ticket.

- Bulk settlements – not all National Rail travel is captured directly in Lennon. National Rail travel on TfL sold Travelcards as well as PTE sold regional travelcards are not directly reported in Lennon and instead are "bulked settled" into Lennon. This also

applies for concessionary travel, for example the Freedom Pass in London where a settlement is entered into Lennon on a quarterly basis with a fixed number of journeys.

- Refunds – refunds in Lennon are not linked to the original ticket sale so are in isolation of each other. There is usually therefore a difference in date of travel of original ticket and refund date, and so a single ticket refunded today in Lennon (-1 journey) will be removed from today's total, but not from the original date of travel for the ticket which most likely would have been in past. This created a particular issue at the start of the pandemic in April 2020 when Lennon was assuming season ticket holders were travelling (so was drip feeding journeys into the system) but in reality, they were not and their ticket was just in the process of being refunded.

- Be aware that season tickets can be an issue on LENNON. They are dumped on the system in one go and so can cause a big spike in ticket distributions. Advise taking a long time period is long enough to manage this.