ARUP

Network Rail and Office of Rail and Road

Independent Reporter

#35532 Train Performance Measures

Reference: 299098-00/001

Issue 2 | 02 April 2024

Job number 299098-00

Ove Arup & Partners Limited
8 Fitzroy Street
London
W1T 4BJ
United Kingdom
arup.com

ARUP

Contents

Tables

Figures

Appendices

Executive Summary

Arup and Winder Phillips Associates (WPA) have been engaged by the Office of Rail and Road (ORR) and Network Rail (NR) to undertake a review of the reliability and accuracy of the reported passenger train performance success measures. This review covers the following train performance metrics at a national, regional and train operator level:

- Cancellations.
- On Time (and associated measures – Time to 3 and Time to 15).

A review of the system reliability and data accuracy for these measures was last completed in 2017. This review enables a re-assessment of the accuracy and reliability of these measures, with particular focus on regional measures and any changes which have been implemented in the meantime. As outlined in the SoW and confirmed in the inception meeting with ORR and NR, this review excluded:

- an assessment of the quality of data within TRUST.
- any focus on pre-plan cancellations or p-coding.

This review, unlike in 2017, has assumed that the data within TRUST is accurate, and so has not covered the TRUST data capture processes, such as berthing offsets. As a result, the confidence scores may not be directly comparable with those from the 2017 review since this did consider the quality of data within TRUST.

Cancellations Measure

NR currently produces the National and Train Operator Cancellations measures using the existing 'PPM Failures' PSS universe. The Regional Cancellations measure will be new for CP7 and as such the development and definition of this measure is not yet fully implemented. NR have provided draft Definitions for the Regional Cancellations measure alongside 'Beta' figures for the measure which is not yet formally published periodically. The Regional Cancellations measure is produced from the new 'On-Time' PSS universe developed as part of the new methodology implemented in 2022. It was noted that NR plan for the National and Train Operator Cancellations measures to be derived from the 'On-Time' universe in CP7. However, this review for these measures has focused on the current processes.

System Reliability – National and Train Operator Cancellations Measure

The Cancellations measures are based solely on data from TRUST, which is fed into PSS. Data is then extracted from the 'PPM Failures' universe within PSS using standardised queries which ensures consistency in the process. The extract data is fed into an Excel spreadsheet, which requires some manual processing. We've noted that this presents a potential risk of error due to requiring, for example, copying down of data, copy/paste of data and use of direct-reference formulas. NR has advised the Reporter Team that plans are in place to automate these processes in the coming months. NR use Taskmaster as a system to support the production process. All activities required to produce the measures each period is captured within the system. Each step in Taskmaster includes a process note, designed to provide clear step-by-step guidance for anyone to potentially complete, providing resilience. Finally, Taskmaster includes a sign-off process for activities which provides an audit trail. NR have a team of four who are fully versed in the production process and aim to cycle around the team as to who is involved in the production to maintain competency. We observe that there appears to be a well-structured process which provides clarity on progress of core activities required to produce the measures, alongside clear guidance documentation. In our view, this supports the reliability, quality, and consistency of the production of the reported measures.

System Reliability – Regional Cancellations Measure

The process for calculating Regional Cancellations was still being developed at the time of this review. NR provided the Reporter Team with a draft document ("Regional Cancellations in BOPSS") which summarises the new set of calculated fields in the BOPSS 'On Time' Universe which will enable this measure to be calculated, along with a set of outputs for review.

From the documentation we have seen, we expect data to be extracted from PSS using standardised queries in line with other measures, which will ensure consistency of calculation. We also expect that similar processes will be put in place as for National and Train Operator Cancellations to transform this data into the published figures, as outlined above.

Data Accuracy – National and Train Operator Cancellations Measure

To review the accuracy of the Cancellations Measure, we have undertaken two core tests:

- Test 1: Validate the filtering and extraction process used by NR. This is to ensure that all valid records are being captured, and no erroneous records are feeding into the measures.
- Test 2: Validate the current calculation process to convert PSS data into published Cancellations figures.

For the purposes of conducting Test 1, NR provided unfiltered data from PSS for all timing records for a sample of 5 days. NR also provided a file summarising all Applicable Trains and whether they were deemed cancellations, along with Daily Flash Reports for direct comparison with reported figures. For Test 1 we completed filtering of the unfiltered data to ensure NR's filtering process is working as intended. Minor anomalies were observed during this review, mostly relating to data within TRUST itself, rather than with the filtering process. Noting the focus of this review did not include the accuracy of data within TRUST. For the sample of 5 days, we were able to replicate the reported daily figures to within 0.1% for both the national and Train Operator Cancellations measures.

For Test 2, we validated that data extracted from PSS was completed correctly using a number of tests. we are confident that the NR calculation process is working as intended and the reported figures are a correct representation of the data extracted from PSS. We have made a few observations as suggested advisory actions to consider improvements in the spreadsheet processing of data, to reduce risk of potential errors.

Data Accuracy – Regional Cancellations Measure

As the Regional Cancellations Measure is not yet published periodically, we have only been able to conduct Test 1 to validate filtering processes completed within PSS prior to extraction from PSS. Using the unfiltered data, we conducted our own independent filtering to compare to the 'Beta' figures reported in the Daily Flash Reports produced by Network Rail. As with the National and Train Operator Cancellations measures, we've been able to reproduce the daily figures for regional cancellations to within 0.1% of the reported figures by Network Rail. We noted one observation with the data filtering process that would benefit from being clarified in the Definitions of Railway Performance document;

- Clarify the method for handing of potential limited number of services that do not have clear indication of cancellations at stop level.

On Time Measure

In March 2022, the approach to calculating On Time was updated to ensure trains which failed to stop at a scheduled station stop were excluded. For such events, TRUST will automatically record an 'actual recorded date time' against the Arrival record, which would reflect the time the train passed the station. Given the train would also have a 'planned date time' for the scheduled stop, such lateness records were erroneously being included in the On Time measure calculation.

The March 2022 update involves the creation of a new 'On Time' Universe, which merges data on train running with that from the Attributions Universe. This enables any timing records where a Cancellation Event was recorded in TRUST for a train and location to be identified.

System Reliability – On Time Measure

The Performance Reporting Manual describes quality assurance processes, and examples have been provided by NR. This includes for each period, a sample export of data from TRUST and PSS which is cross compared. The purpose is to ensure that the delay minutes and cancellations reported in PSS match what was input into TRUST. Any differences are thoroughly checked, and valid reasons are documented in reports. NR has also recently implemented an assurance check to verify the accuracy of stations used for the On time

measure. As with the Cancellations measures, we observe that NR appears to have adopted a well-structured process, supported by the Taskmaster system, which provides clarity on progress of core activities required to produce the measures, alongside clear guidance documentation. Checks are put in place to assure the data that is being reported. In our view, this supports the reliability, quality and consistency of the production of the reported measures.

Data Accuracy – On Time Measure

In the same way as outlined with the Cancellations measure, we have conducted two core tests, Test 1 to validate the filtering of data, and Test 2 to validate the calculation process to convert PSS data into published Cancellations.

For Test 1, we have replicated the filtering process for the On Time measure based on the same sample of 5 days of unfiltered PSS data. For all dates we were able to directly match the number of records being used to those in the published statistics. We were also able to directly match the calculated On Time, Time to 3 and Time to 15 measures for these 5 days, so providing confidence in the accuracy of the reported figures.

For Test 2, we validated that data extracted from PSS was completed correctly using a number of checks. We are confident that the NR calculation process is working as intended and the reported figures are a correct representation of the data extracted from PSS. As with the Cancellations measure, we have made a few observations as suggested advisory actions to consider improvements in the spreadsheet processing of data.

Conclusions

The following Table 1 outlines a summary of the confidence grades for each of the measures considered in this review based upon the analysis outlined in this report. Note that the scoring under this mandate cannot directly be compared to the review completed in 2017, with the difference in scope where TRUST data is taken as-is for this review.

Table 1 – Summary of confidence grades

Measure	Current Confidence Grade	Commentary
National and TOC Cancellations Measure	A1*	Our review of NR documentation, data and processes has not identified any material concerns with the reliability of the reported National and Operator Cancellations measure. We have also been able to replicate the reported figures to within 0.1% giving an accuracy score of 1*.
On Time Measure (and Associated Time to 3 and Time to 15 Measures)	A1*	Our review of NR documentation, data and processes has not identified any material concerns with the reliability of the reported National and Operator Cancellations measure. We have also been able to replicate the reported figures to within 0.1% giving an accuracy score of 1*.

As the Regional Cancellations measure is not yet formally produced and as part of the review we have therefore not been able to conduct all tests, the following Table 2 presents provisional scores on the assumption that the process for reporting the Regional Cancellations builds on existing process as described in our review. We would also expect NR to complete its planned update for the Definition documentation.

Table 2 – Provisional confidence grades

Measure	Current Confidence Grade	Commentary
Regional Cancellations Measure	A1* *Provisional*	We do not expect there to likely be any material concerns with the reliability of the Regional Cancellations measure on the assumption that the process builds on the NR documentation, and processes reviewed for existing measures. We have also been able to replicate the reported figures to within 0.1% giving a provisional accuracy score of 1*. As the Regional Cancellations measure is not currently reported, it has not been possible to conduct Test 2 for this measure.

We do not have any specific recommendations to be made under the scope of this mandate., However we have included 4 suggestions for potential improvements to be considered. Three relating to improvements to spreadsheet processing of data and the remaining one to provide greater clarification in the Definitions of Railway Performance Metrics on specific cases.

Acknowledgements

The Independent Reporter Team would like to thank both ORR and Network Rail staff for their assistance with this study.

1. Introduction

1.1 Background

Arup and Winder Phillips Associates (WPA) have been engaged by the Office of Rail and Road (ORR) and Network Rail (NR) to undertake a review of the reliability and accuracy of the reported passenger train performance success measures which NR will be monitored against during Control Period 7 (CP7). This project is undertaken by Arup in its role as Independent Reporter.

This review covers the following train performance metrics at a national, regional and train operator level:

- On Time (and associated measures – Time to 3 and Time to 15)
- Cancellations

The scope of this study was defined in the Statement of Work (SoW) #37788, a copy of which is included in Appendix A.1.

1.2 Mandate Aims and Requirements

The objective of this review is to measure the system reliability and data accuracy of the reported On Time (and associated measures – Time to 3 and Time to 15) and Cancellations measures. The SoW outlines that, as part of the important role in monitoring train performance and being success measures in CP7, it is critical that NR, ORR and rail industry stakeholders have assurance of the quality of the data and robustness of these measures.

Arup and WPA conducted a review of the system reliability and data accuracy for these measures in 2017 as part of a wider review of the new performance metrics being introduced for Control Period 6 (CP6)[1]. This review enabled a re-assessment of the accuracy and reliability of these measures, with particular focus on regional measures and any changes which have been implemented in the meantime. In particular, the SoW identified a change applied to the calculation of the On Time measure in March 2022 to better reflect trains which failed to stop at one or more scheduled stops.

The scope of this review covers three areas:

- Governance for collecting the data for the On Time (and associated measures – Time to 3 and Time to 15) and Cancellations measures from TRUST and the improved methodology (in March 2022) for transforming it into outputs provided across the rail industry.
- Reliability, quality, consistency, completeness and accuracy of reported data for On Time (and associated measures – Time to 3 and Time to 15) and Cancellations measures at a national, regional and train operator level.
- Processes to produce, quality assure and provide consistent period end data to customer/stakeholders including ORR.

As part of this review, we provided a confidence grading for each measure, based on an alpha (system reliability) and numeric (data accuracy) grading based on the most up to date dataset available (up to Period 10 of 2023/24). The grading system for the confidence rating is shown in Appendix A.2.

As outlined in the SoW and confirmed in the inception meeting with ORR and NR, this review excluded:

- an assessment of the quality of data within TRUST
- any focus on pre-plan cancellations or p-coding.

This review focused on the reliability and accuracy of the conversion of data held within TRUST into the reported metrics. This review, unlike in 2017, has assumed that the data within TRUST is accurate, and so

[1] The 2017 ‘Review of New Performance Metrics’ can be accessed via https://www.orr.gov.uk/media/16225

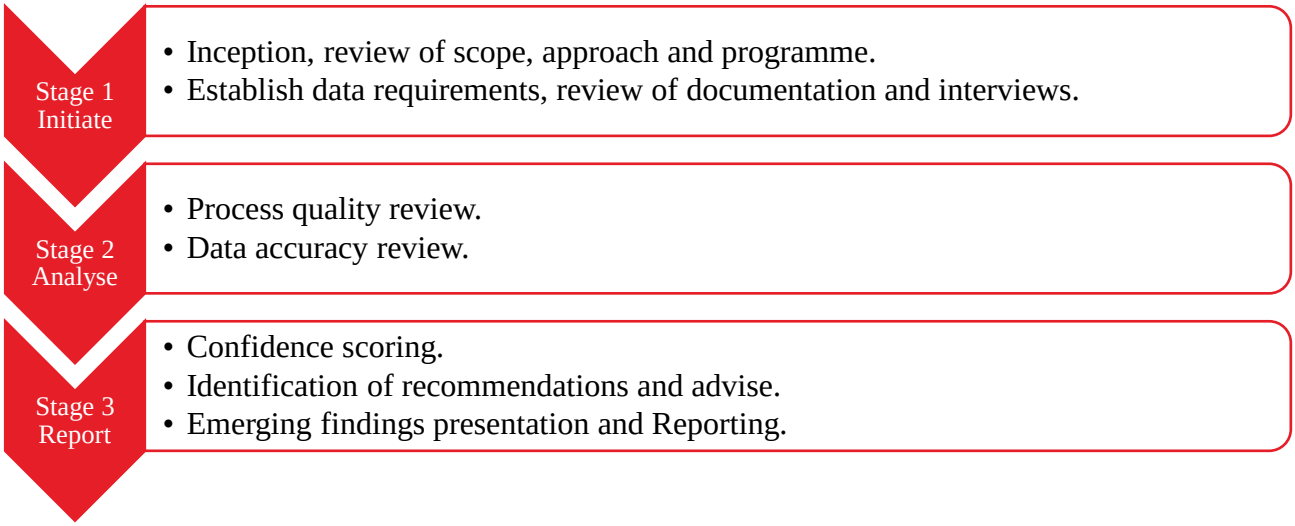
has not covered the TRUST data capture processes, such as berthing offsets. As a result, the confidence scores may not be directly comparable with those from the 2017 review, since this did consider the quality of data within TRUST.

The processes for producing the Regional Cancellations measure were still in development at the time of this review. As such, we have reviewed the latest position with this measure and provided our emerging views on this process, recognising that further work by NR is still in progress.

1.3 Our Approach

The approach that we adopted for this study was designed to provide an assessment of NR's reporting process, procedures and governance, alongside an audit of the underlying data calculation approach to review the accuracy of reported results. Our approach is summarised in Figure 1.

Figure 1 – Review Approach



During the engagement phase, we met with representatives from the NR Network Performance Team to gain a full understanding of the current process for calculating the measures, any changes that are planned ahead of CP7, and to agree a specification for the data that would be provided to support this study.

To support our review, NR has provided a set of documents outlining the processes undertaken to calculate the measures and quality assurance tasks undertaken. A full list of documents provided is covered in Appendix A.3.

Sample data from PSS has also been provided to support our review of the accuracy of the reported measures.

Our data validation approach has involved effectively replicating the process adopted by NR to calculate the measures based on data held within PSS. We have undertaken these tests independently using GitHub and Python scripts to enable tests to be consistent, repeatable and clear. Our tests covered three core areas:

- Validate the current calculation process to convert PSS data into published On Time and Cancellations figures.
- Validate the filtering process used by NR to ensure that all valid records are being captured, and no erroneous records are feeding into the measures.
- Verify that assurance has been undertaken following a change in system architecture within PSS in 2022, designed to improve both the accuracy and speed of metric production.

2. Review of Cancellations Measure

2.1 Overview

This section summarises the findings from our review of the system reliability and data accuracy related to the Cancellations measure. A description of the measure is provided, followed by sections outlining the following areas which cover the questions in the Mandate:

- Approach for converting TRUST/PSS data into the Cancellations measure;
- System reliability of the calculation process (covering reliability, quality and consistency); and
- Data accuracy for the measure (covering accuracy and completeness).

A confidence grading is provided at the end of this section based on the findings of our review.

2.2 Definition of Cancellations Measure

2.2.1 Definition of National Cancellations

The Cancellations measure is officially defined within the "**Definitions of Railway Performance Metrics_v3.03**" document (dated September 2020).

The one-line description of Cancellations is:

"The percentage of planned trains which either did not run their full planned journey or did not call at all their planned station stops[2]"

A train is classed as fully cancelled (weighted cancellation = 1) if it runs less than 50% of its planned journey distance, including trains which did not run at all. It is classed as part cancelled (weighted cancellation = 0.5) if it runs between 50% (inclusive) and 100% (exclusive) of its journey distance or if it completes its full journey length but fails to stop at one or more of the stations at which it is planned to call (termed 'fail to stop').

The measure is calculated as the total number of weighted cancellations as a percentage of the total number of planned trains in the Applicable Timetable, i.e.

Cancellations Measure = Sum of Weighted Cancellations / Sum of Planned Trains

The National Cancellations measure is therefore measured against all trains in the Applicable Timetable[3] which are operated by a government funded operator (including MerseyRail and Rail for London concessions[4]) or by one of four Open Access Operators which are deemed by the ORR to operate a core passenger service as part of the UK national rail offering (Hull Trains, Grand Central, Heathrow Express and Lumo).

The Railway Performance Definitions document clarifies that information on cancellations and fail-to-stops is to be entered into the TRUST system on the day of the event through use of defined TRUST 'Event Codes'[5]. If a cancellation event was not entered into TRUST on the day, then it is the responsibility of Train Operating Companies to advise NR in time to be included in the data systems. If this information is received after 'Day 8' (i.e. 8 days after the event, when TRUST can no longer be edited), then this information would need to be updated within PSS. Such a change is required to be signed off by the Route Performance

[2] Planned Station Stops always include the planned origin and terminating stations of the train.

[3] The passenger timetable which reflects the working timetable for the Passenger Services required to be drawn up by NR in accordance with the Track Access Conditions, as at 22:00 on the immediately preceding day.

[4] This includes services which run wholly off NR infrastructure, e.g. Elizabeth Line services between Paddington and Abbey Wood.

[5] TRUST Cancellation codes are C= Full Cancellation; P= Did not reach destination; O= Change of origin; D= Diverted Service; F= Failed to Stop.

Measurement Manager and validated by the Network Performance Team which ensures governance over data amendments within PSS.

2.2.2 Definition of Operator Cancellations

The definition of the Cancellations Measure can be directly translated into a Train Operator measure of cancellations based on those trains in the Applicable Timetable which are planned to be run by each operator.

2.2.3 Definition of Regional Cancellations

The Regional Cancellations measure will be new for CP7 and as such, the definition of this metric is not yet included in v3.03 of the Definitions of Railway Performance Metrics document. However, NR has provided the Reporter Team with proposed wording for this measure which was agreed by the Performance Metric Steering Group (PMSG) at the end of January. It is intended that this will be included in the next release of the Definitions of Railway Performance Metrics document.

At the high level, Regional Cancellations are described as:

> *"The percentage of planned trains which either did not run their full planned journey or did not call at all their planned station stops within a geographic region"*

To support this definition, NR has also provided the Reporter Team with a draft document which defines the planned approach for measuring regional cancellations ("**Regional Cancellations Definition**") in more detail. This document includes illustrative examples.

This document further clarifies that "*The measure is essentially taking the national figure and apportioning the failures between the Regions based on where the train failed to call*." The main premise of this is to ensure that the National and Train Operator cancellations figures would be unaltered if the Regional Cancellations were summed up."

To calculate this measure, each planned train is apportioned between Regions[6] based on the proportion of scheduled stops within each Region, for example if 3 out of 10 planned station stops were in Region A, then 30% of the train is apportioned to Region A. Similarly, weighted cancellations (termed 'Cancellation Value[7]') are apportioned based on the proportion of missed stops occurring in each Region.

The example below (Table 3) is taken from the draft definition document provided by NR illustrate the calculation of Apportioned Trains and Apportioned Cancellations Values.

[6] This includes the 5 NR Regions plus any stops off the NR geography (e.g. TfL network) as a separate Region.

[7] Weighted in line with standard definition, so the Cancellation Value for a Full Cancellation is 1.0 and the Cancellation Value for a Part Cancellation is 0.5.

Table 3 – NR illustration of the calculation of Apportioned Trains and Apportioned Cancellation Values

Train	Planned Stops (Region X)	Planned Stops (National)	Apportioned Trains (Region X)	Cancellation Value	Missed Stops (Region X)	Missed Stops (National)	Apportioned Cancellations (Region X)
A	5	5	5 / 5 = 1.0	1.0	5	5	1.0 * 5 / 5 = 1.0
B	4	5	4 / 5 = 0.8	1.0	4	5	1.0 * 4 / 5 = 0.8
C	3	5	3 / 5 = 0.6	1.0	0	3	1.0 * 0 / 3 = 0.0
D	3	5	3 / 5 = 0.6	0.5	0	2	0.5 * 0 / 2 = 0.0
E	3	5	3 / 5 = 0.6	0.5	1	1	0.5 * 1 / 1 = 0.5
F	3	5	3 / 5 = 0.6	0.5	1	2	0.5 * 1 / 2 = 0.25
G	1	5	1 / 5 = 0.2	0.5	0	1	0.5 * 0 / 1 = 0.0
TOTAL			**4.4**				**2.55**

The number of apportioned trains and apportioned cancellations are then summed for each Region. For example, using the figures above the number of apportioned trains for Region X would be 4.4 and the number of apportioned cancellations for Region X would be 2.55.

From this, the regional cancellations measure can be calculated as:

Regional Cancellations = Sum of Apportioned Cancellations Value / Sum of Apportioned Trains

2.3 System Reliability: Cancellations Measure

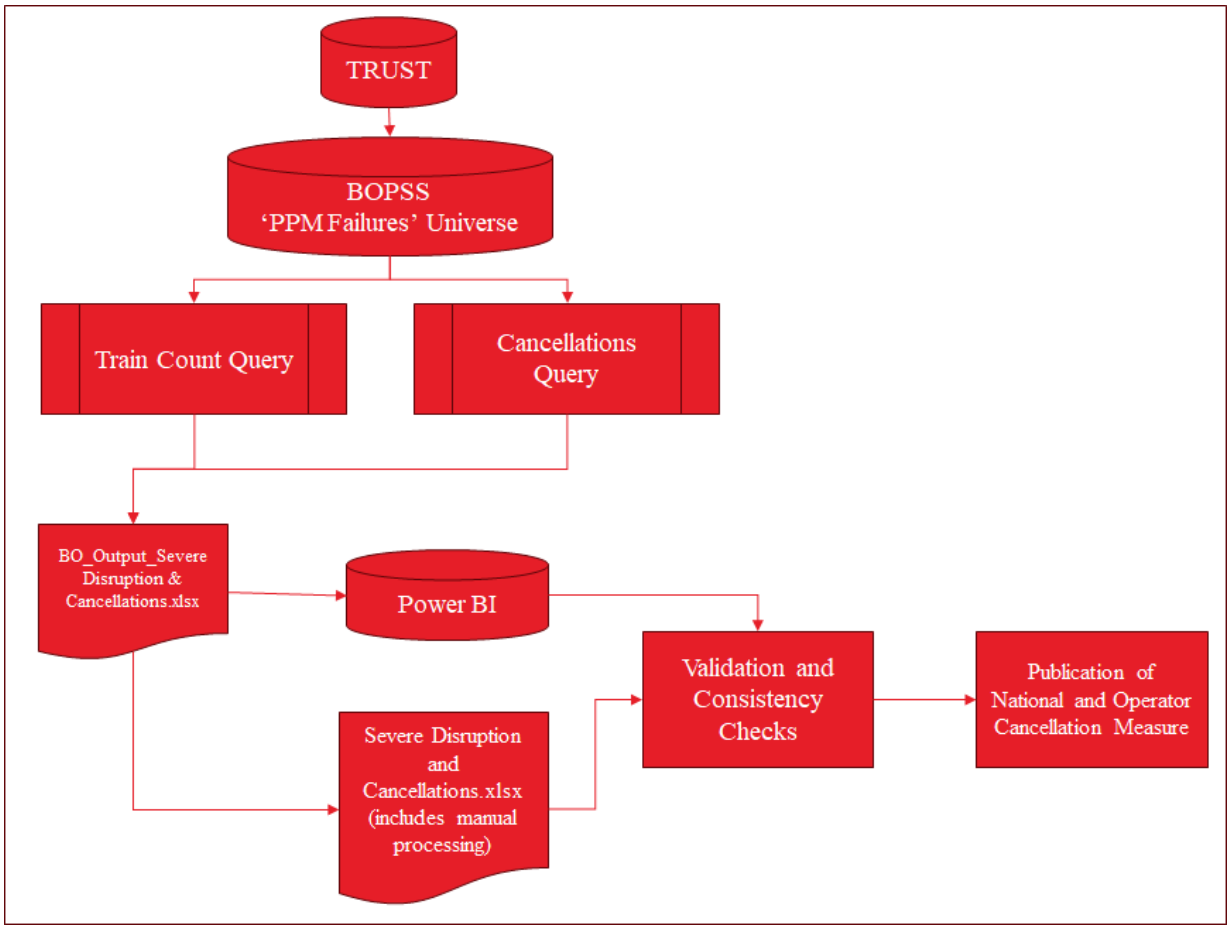
2.3.1 National and Operator Cancellations

Two documents have been provided to the Reporter Team which outline the process for producing the National and Operator Cancellations measures each period:

- "**Cancellations**": describes the Business Objects (BOPSS) queries for extracting data from PSS, dated 1st April 2023
- "**NR-NPRT-WI-004 (Severe Disruption and Cancellations)**": describes a step-by-step guide for updating the data files with the data from the BOPSS queries including links to relevant files, dated 8th March 2022.

We have summarised this process in Figure 2.

Figure 2 – Process for producing the National and Operator Cancellations



The Cancellations Measure is based solely on data within TRUST which is fed into PSS. Data is extracted from the "PPM Failures" Universe in PSS using two Business Objects queries which filter only on train services which are relevant to the measure, so those trains:

- with an applicable flag, i.e. scheduled to run in passenger service (i.e. revenue earning) as at 22:00 the day before operation; and
- which are run by a government funded operator or by one of four Open Access Operators as described in Section 2.2.1.

The use of such standardised queries ensures that data is extracted in a consistent manner each period, including historical data for refresh purposes. This data is then uploaded into a spreadsheet from which the National and Train Operator Cancellations Measures are calculated. Some simplistic manual processing is undertaken at this stage, primarily copying and pasting data between files, and then copying formulae down in certain worksheets to update for latest periods. Further observations on this process are covered in Section 2.4. NR has advised the Reporter Team that plans are in place to automate these processes in the coming months.

The latest data on the Cancellations Measure is stored on Network Rail's Performance SharePoint site which can be accessed by industry performance practitioners. This SharePoint site also provides Power-BI dashboards BI which enables users to visualise the data.

NR advised that the current data refresh policy is that the data is refreshed on day 2 of the period as a provisional outcome and confirmed on day 8. It is then fully refreshed at the end of the financial year. There is a comment in the "Cancellations" document that this will be amended in CP7, such that the data will be refreshed twice after each period (and confirmed in discussion with the NR NPT team):

- Day 2 following period end for provisional results;
- Day 8 following period end for final results, to be provided to ORR.

NR NPT advised that all data for all periods in the existing Control Period is refreshed every period, and so this ensures that any subsequent changes to the underlying PSS data will be picked up through this process and fed into the reported figures.

A further document that has recently been produced by NR is the Performance Reporting Manual, which forms a module of the wider Performance Measurement Manual. Version 1 of this document (November 2023) was provided to the Reporter team for review ("**29 – Metrics Reporting Manual_V1**"). We were advised that the purpose of this document is to fill a gap between the official performance measures definitions document and the detailed process documents which provide step-by-step guidance on producing the measures.

This document outlines the key principles and definitions of the measures currently being reported, along with governance and assurance activities which are required to be followed to ensure consistency and accuracy of reporting. As such it provides more detailed guidance to the system users and can be adjusted without necessarily having to revert minor changes to PMSG providing they remain in line with the officially agreed definitions.

The Performance Reporting Manual describes quality assurance activities undertaken on the Cancellations measures, which focus around having a scheduled and documented programme of work to ensure the required data is updated and completed within required timescales and subject to periodic testing.

To support this production process, all activities required to be run each period are captured in NR's Taskmaster system. Each activity includes a 'process note', which is designed to provide clear step-by-step guidance on how to complete that activity. NR advised that these notes are designed so that anyone could pick up and run the activity, thus providing resilience. Once an activity has been completed, it must be signed off on the system so providing an audit trail of who has undertaken each activity and when.

Within Taskmaster, NR have introduced a set of checks to run to ensure that all files have been updated and correctly uploaded to SharePoint each period using a file synchroniser process.

As part of the quality assurance process, each period a sample export of data from TRUST and PSS is cross compared. The purpose of this is to assure that the delay minutes and cancellations which appear in PSS match what was input into TRUST (so this check is relevant to both Cancellations and On Time measures). Four examples were provided to the Reporter Team looking at specific TOCs and stations (Avanti West Coast at Euston, EMR at St Pancras, GWR at Paddington and CrossCountry at Penzance). This highlights that where any differences are spotted, the team have checked and confirmed that the reason is valid (e.g. linked to a non-applicable train). This is documented in the reports. This provides confidence that relevant data is not being lost/amended in the transition from TRUST to PSS.

There is a team of 4 people who are fully versed in the production schedule. The team aims to cycle round who is involved to keep competency up. We were also advised that there are other people within the wider performance team who have previously undertaken the production process and so could pick up and run if necessary. In our view, this provides a good level of resilience to the process, which is supported by the enhanced step-by-step documentation (compared with the 2017 review), ensuring the reliability of the process for producing these measures each period.

The NR NPT team advised that an independent review of the periodic performance measures is undertaken by a member of the NPT team not involved in production before releasing to ORR. This is designed to ensure a 'final pair of eyes' on the figures to supplement other checks that are undertaken.

We observe that there appears to be a well-structured process, supported by the Taskmaster system, which provides clarity on progress of core activities required to produce the measures, alongside clear guidance documentation. In our view, this supports the reliability, quality, and consistency of the production of the reported measures.

From discussion with the NR NPT team, we understand that, while the cancellations measure is currently calculated from the 'PPM Failures' Universe within BOPSS, there is a plan to transition to being calculated from the 'On Time' Universe ahead the start of CP7. This is a new Universe in BOPSS which was created in 2022 to facilitate the improved calculation of On Time by excluding Fail to Stop records (see Section 3.2.2 for more detail) and is also required to support the Regional Cancellations Measure calculation. Extracting data related to National and Train Operator cancellations from this same Universe will ensure all

Cancellations measures are being calculated from the same data source. This new Universe also improves the efficiency of data processing within PSS which will facilitate the planned move to refreshing data twice at the end of each period, as outlined above and bringing this in line with the current refresh policy for the On Time measure.

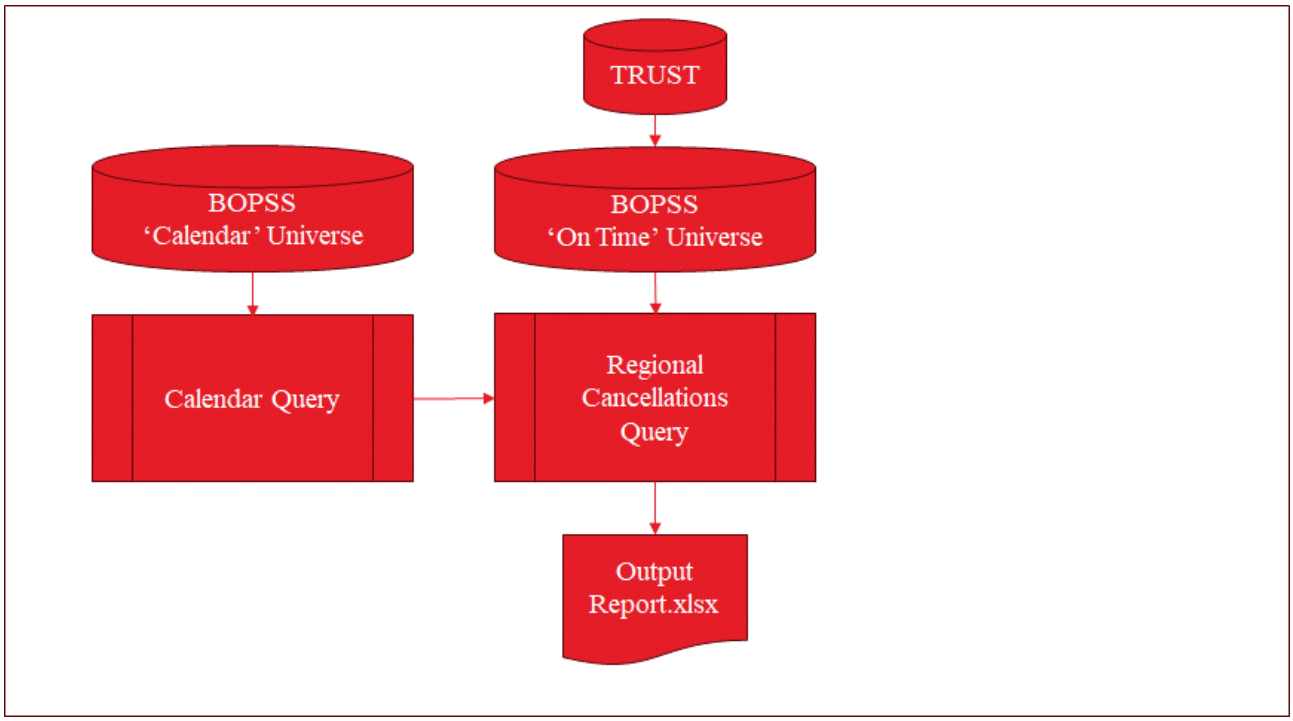
Our review has been based on documents which relate to the current process for National and Train Operator Cancellations. NR has advised that all process documentation and guidance will be updated ahead of CP7 to reflect this revision. Therefore, the process chart at the start of this section will need to be updated to reflect the new BOPSS Universe and queries.

2.3.2 Regional Cancellations Measure

The process for calculating Regional Cancellations was still being developed at the time of this review. NR has provided the Reporter Team with a draft document ("**Regional Cancellations in BOPSS**") which summarises the new set of calculated fields in the BOPSS 'On Time' Universe which will enable this measure to be calculated, along with a set of outputs for review.

We have summarised the draft process for calculating this measure in Figure 3. This measure is not yet being published, so the process for transition of this data from the 'Output Report' file through to publication to ORR is not yet finalised. As such, this process chart is not yet complete.

Figure 3 – Draft process for producing the Regional Cancellations



The draft documentation provided to describe the query used in the On Time Universe has merged a number of previously separate filters (PfPI Trains, Applicable Train Flag) into a single filter 'Applicable Train' which enables more efficient processing of the data.

From the documentation we have seen, we expect data to be extracted from PSS using standardised queries in line with other measures, which will ensure consistency of calculation. We also expect that similar processes will be put in place as for National and Train Operator Cancellations to transform this data into the published figures, as outlined above.

2.4 Data Accuracy: Cancellations Measure

To review the accuracy of the Cancellations Measure, we have undertaken two core tests:

- Test 1: Validate the filtering and extraction process used by NR. This is to ensure that all valid records are being captured, and no erroneous records are feeding into the measures.

- Test 2: Validate the current calculation process to convert PSS data into published Cancellations figures.

The following presents the outcomes of the review for National and TOC measures for Test 1 and Test 2, followed by the outcomes of Test 1 for the Regional Cancellations measure. As the Regional Cancellations measure is not currently reported, it has not been possible to conduct Test 2 for this measure.

2.4.1 Test 1: Validation of Extraction of PSS Data – National and TOC Cancellations Measure.

Test 1 is focused on validating the data from TRUST to PSS prior to post-processing which is considered in Test 2. In effect, it is sampling the filtering processes completed within PSS prior to extraction from PSS.

NR provided unfiltered data from PSS containing all timing records for a sample of 5 recent days so we could validate the trains which were included in the calculation of the Cancellations measure. NR also provided the Reporter Team with a file summarising all Applicable Trains on those days and whether they were deemed a cancellation, along with Daily Flash Reports to enable a direct comparison with reported figures. From this, it allowed us to check whether we could effectively calculate the measures from first principles.

The purpose of this check is to ensure that NR's filtering process is working as intended. Given the queries are standardised, we agreed with NR that 5 days covering the full network should be a sufficient sample to provide confidence in this process. We also noted that it was a time-consuming process for NR to provide data at this level of granularity, and so this was also a pragmatic decision. The selected 5 days included a mix of 'good' and 'bad' performance days, covering both weekdays and weekends.

For each sample day, we replicated the filters that are applied for cancellations. This enabled us to review and confirm records that were excluded and check whether any records have been included erroneously. As part of the review, we queried some potential anomalies within the data, and whilst minor in the wider context of the measure, mostly were a result of issues with data within TRUST itself and not as a result of filtering data, noting that the focus of this review does not include accuracy of data within TRUST. We also observed that some services whilst operating as a single train in the timetable are presented as two separate services in the data. This applies to GTR services operating through the Thameslink core and Elizabeth Line services. For the cancellations measures this means these services are in effect reported twice, with regional measures having stops in each service apportioned to their respective region for portion of service. This is the agreed approach for cross-London services as defined in Appendix B of the Definitions of Railway Performance document.

For the sample of 5 days, we were able to replicate the reported daily figures to within 0.1% for both the national and Train Operator Cancellations measures, as illustrated in Table 4.

Table 4 – Replicated sample of daily figures for the National and TOC cancellations measures.

	16th June 2023	22nd November 2023	5th December 2023	16th December 2023	21st December 2023
Total trains (Unfiltered dataset)	32,199	32,455	35,301	29,248	32,661
Excluded Records:					
Non passenger Operator	2,389	2,438	2,448	1,548	2,256
Non applicable train	189	161	275	267	409
Non PfPI train	7,239	7,382	8,112	5,771	7,269
Train running on excluded TOPS **code**	12	13	8	6	7
Scheduled cancel	445	426	5,367	1,066	785
Planned part cancellation	0	0	0	0	0
Planned full cancellation	34	11	126	53	23
Total Excluded Trains	10,308	10,431	16,336	8,711	10,749
Comparison of recorded stops: Arup calculations vs NR official data					
Filtered Trains (Arup)	21,891	22,024	18,965	20,537	21,912

Filtered Trains (NR)	21,891	22,024	18,965	20,537	21,912
Difference	0	0	0	0	0

2.4.2 Test 2: Validation of Calculation of National and Train Operator Cancellations Measures from PSS Data

Test 2 reviews the post-processing undertaken to produce the National and TOC Cancellations measures from the data extracted from PSS. In order to conduct this test we have applied the following checks as part of the review:

- Validate that the data extracted from PSS (in file 'BO_Output_Severe Disruption & Cancellations.xlsx') was correctly copied into appropriate locations in the 'Severe Disruption and Cancellations.xlsx' workbook.
- Confirm the validity of the formulae calculations within the 'Severe Disruption and Cancellations.xlsx' workbook.
- Apply consistency checks for formulae contained within the 'Severe Disruption and Cancellations.xlsx' workbook.
- Check that the charts contained within the 'Severe Disruption and Cancellations.xlsx' workbook correctly reference the selected period of data.

From our tests, we are confident that the NR calculation process is working as intended and the reported figures are a correct representation of the data extracted from PSS.

From our review, we have made the following observations on the 'Severe Disruption and Cancellations.xlsx' workbook which NR may like to consider where potential risk of error may occur, and so to provide additional resilience to the update process:

- The formulae used to calculate moving annual average figures in the 'TOC Data' sheet are based on direct cell references to previous rows. There is a risk that if the order in which the data is extracted from PSS changed in one period or if the number of Operators for which data is extracted changes, this could result in incorrect calculations. We would propose a generic consistent formula is applied based on the relevant TOC and the Period indices (to identify the relevant periods for the calculation).
- We suggest it would be useful to build some internal validation into the workbook to support validation of the metric, for example a check for duplicate values (e.g. if data for a particular period and TOC is manually pasted into this worksheet twice).
- There were some inconsistencies identified in the formulae in the 'TOC Data' sheet. It is advised that a visual reference (e.g. different coloured columns) be applied to highlight this difference. For example, inconsistent formulae were observed in Column I of the "TOC Data" sheet between rows 2281 and 2283 when the Lumo Operator was first introduced (this links to the first bullet above, where it is suggested that a more generic consistent formula is applied).
- While the Severely Disrupted Days measure is outside scope, we observe that within the 'Severe_National' sheet, the data in columns B and C could be swapped to enable the relevant periods to be copied directly from the input file.

2.4.3 Test 1: Validation of Calculation of Regional Cancellations Metrics

As the Regional Cancellations Measure is not yet published periodically, we have only been able to conduct Test 1 to validate filtering processes completed within PSS prior to extraction from PSS. This test uses the same 5-day sample of unfiltered PSS data used for testing National and Train Operator Cancellations measures. Using the unfiltered data, we conducted our own independent filtering to compare to the reported measures in the Daily Flash Reports (labelled as 'Beta' figures to distinguish the process for calculating these as still in testing and development mode) produced by Network Rail. As with the National and TOC Cancellations measures, we've been able to reproduce the daily figures for regional cancellations to within 0.1% of the reported figures by Network Rail.

As part of the process in undertaking Test 1 we made the following observation:

- Although minor in relation to the reported measures (affecting a total of 3 services), within the 5 sample days used, we observed some services with no clear indication of cancellations at stop level, some services were indicated as either full or partial cancelled but excluded "failed to stop" flags for stops within the service. Whilst this relates to data within TRUST itself and is outside the scope of this mandate, following clarification with NR, where such limited circumstances apply the current assumption is to assume that where actual recorded times are present in the train record, these would be assumed as successful stops, and where no recorded times are present assumed as cancelled stops. NR have noted that they are still reviewing the correct handling method for such trains and expect instances of this to be resolved by Day 8. Overall, this is likely to be very minor and does not affect the accuracy scoring, however, we observe the assumption is not explicitly covered in the Definitions of Railway Performance document and may be a useful amendment for clarity.

2.5 Confidence Rating for Cancellations Measure

Our audit has identified no material concerns with the reliability of the reported National and Train Operator Cancellations measures. The use of automated queries to extract data ensures a consistent approach, while the process is well documented with further resilience provided by the size of the team available to run the process.

Reflecting on our review, the National and Train Operator Cancellations Measures are awarded **'A'** for overall reliability. This is an improvement from the 'B' awarded in the 2017 review and reflects the improvements made in the documentation for this measure.

Reflecting on our review of the accuracy and processes undertaken, the Cancellations Measures are awarded a '**1***' for overall accuracy.

It should be noted that this score cannot be directly compared with the score awarded in the 2017 review. The current review focuses solely on the conversion of TRUST data into the reported measures, and so the score reflects our views of the reliability and accuracy of this process. The reliability and accuracy scores from the 2017 review also heavily drew upon issues observed with the capture of data within TRUST, for example reporting of Fail to Stop events in TRUST. TRUST data is outside the remit of this review. If the scope of the 2017 review was similar to the current review, then the scoring is likely to have been more consistent with this review.

On the assumption that the process for reporting Regional Cancellations builds on existing processes as described in our review, we would expect to score this measure on a similar level to the national and operator measures. A provisional score of A1* is expected. As the Regional Cancellations measure is not yet formally produced and as part of the review, we have not been able to conduct all tests, we can only indicate a provisional score.

3. Review of On Time Measure (and Associated Time to 3 and Time to15 Measures)

3.1 Overview

This section summarises the findings from our review of the system reliability and data accuracy related to the On Time measure (and associated Time to 3 and Time to15 measures). The process that NR use to generate each of the On Time, Time to 3 and Time to15 measures is identical. As such, unless otherwise specified, any reference in this section to the On Time measure should be taken to equally apply to the Time to 3 and Time to 15 measures.

A description of the measure is provided, followed by sections outlining the following areas which cover the questions in the Mandate:

- Approach for converting TRUST/PSS data into the On Time measure;
- System reliability of the calculation process (covering reliability, quality and consistency); and
- Data accuracy for the measure (covering accuracy and completeness);

A confidence grading is provided at the end of this section based on the findings of our review.

3.2 Definition of On Time Measure

3.2.1 Definition

The On Time measure is officially defined within the "**Definitions of Railway Performance Metrics_v3.03**" document (dated September 2020) alongside the wider group of punctuality measures.

Punctuality is defined as being "*measured through assessing the lateness experienced at each recorded station stop*" where a recorded station stop is a "*location in the applicable timetable with both a planned date time (GBTT)* ***and*** *an actual recorded date time (according to TRUST).*[8]"

On Time punctuality is defined as

"The percentage of Recorded Station Stops called at on time or early"

A train is defined as On Time at a station where the lateness is recorded as less than 1 minute (up to and including 59 seconds late) as measured against the Public Timetable time (GBTT) within TRUST.

All recorded station stops are directly linked to an Train Operator (via the train schedule) or a Region (via a mapping of station to Region). Therefore, Train Operator and Regional On Time measures are calculated based on filtering on relevant records for that Train Operator or Region respectively.

For completeness, the associated measures, Time to 3 and Time to 15 are defined in the same way, with different lateness thresholds applied, as follows:

- Time to 3: The percentage of Recorded Station Stops called at within 3 minutes (up to and including 2 minutes, 59 seconds late) of the planned time.
- Time to 15: The percentage of Recorded Station Stops called at within 15 minutes (up to and including 14 minutes, 59 seconds late) of the planned time.

[8] Stations which are pick-up only or drop-down only are included within the measure because there are planned and actual date times within TRUST for them. However, request stops are only included when they are activated and an actual arrival time is recorded.

3.2.2 March 2022 Update

In March 2022, the approach to calculating On Time was updated to ensure trains which failed to stop at a scheduled station stop were excluded.

For such events, TRUST will automatically record an 'actual recorded date time' against the Arrival record, which would reflect the time the train passed the station. Given the train would also have a 'planned date time' for the scheduled stop, such lateness records were erroneously being included in the On Time measure calculation.

In such circumstance, there is reliance on the Operator or Control being aware of the failure to stop at a scheduled station call and manually entering a Fail to Stop ("F") cancellation event in TRUST against that train and station stop. Once this is done, this triggers a TRUST Delay Attribution (TDA) alert, so enabling the cancellation event to be attributed to a cause. This information appears in the Attribution Universe in PSS.

The March 2022 update involves the creation of a new 'On Time' Universe, which merges data on train running with that from the Attributions Universe. This enables any timing records where a Cancellation Event was recorded in TRUST for a train and location to be identified.

Using this new dataset, 'Applicable Stops' can be defined as "*location in the applicable timetable with both a planned date time (GBTT)* ***and*** *an actual recorded date time (according to TRUST)* ***and where a 'Fail to Stop' cancellation event was not reported in TRUST****".*

As such, only records for both Applicable Trains and Applicable Stops are included in the calculation of the punctuality measures.

3.3 System Reliability: On Time Measure

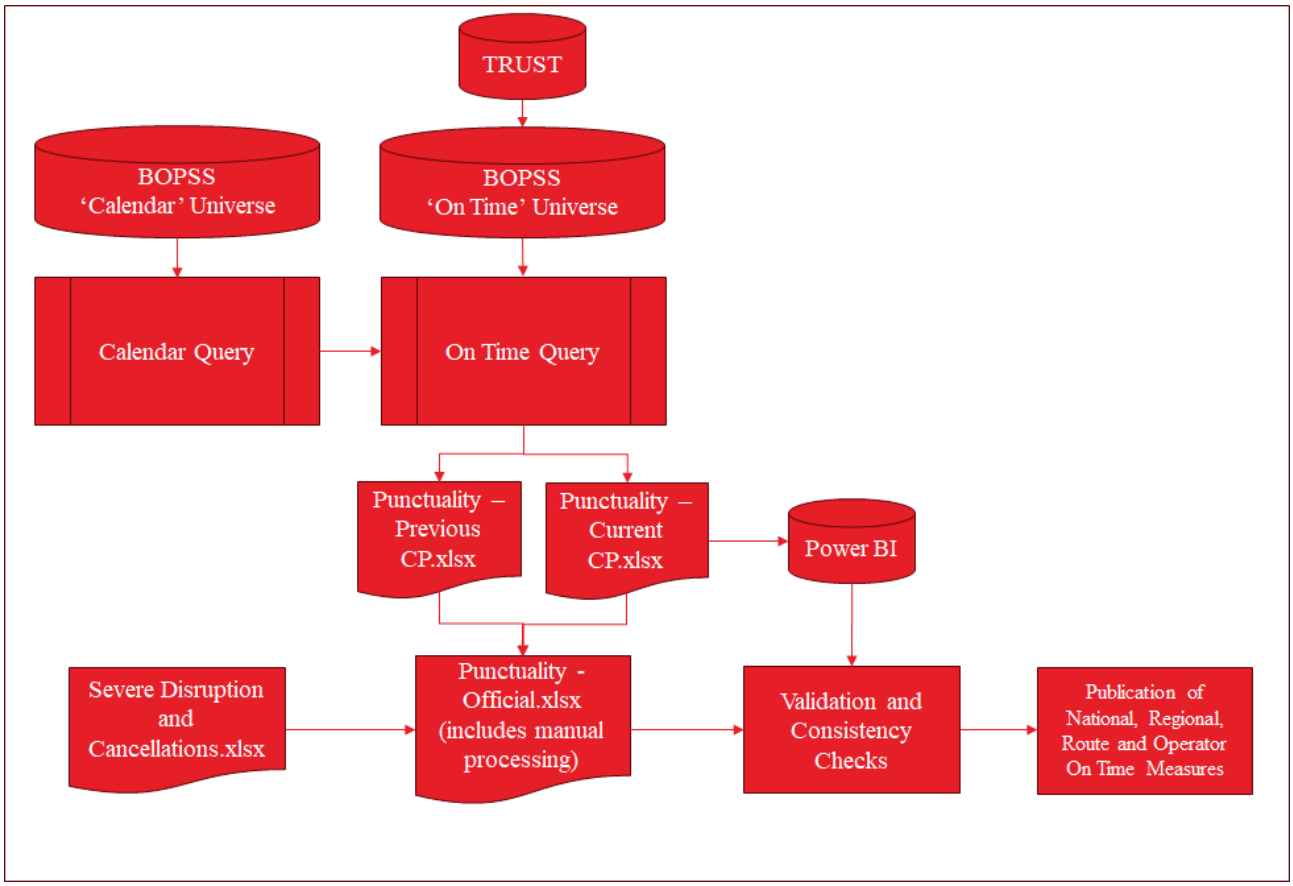
Two documents have been provided to the Independent Reporter team which outline the process that is required to be followed to produce the On Time (and associated punctuality measures) each period.

The BOPSS queries for extracting data for calculating the Punctuality Measures are described in document "**Punctuality**" (titled 'NR-NPT-Metric Definition Punctuality Refresh Schedule and Query Definition', dated 1 April 2023). This clearly outlines the filters that are applied to the PSS data.

The process to update relevant files with the refreshed data from the BOPSS queries is outlined in document "**NR-NROT-WI-010 (Punctuality – Official)**", dated 18th October 2023 (version 2.1). This document provides links to the core files required to run this update, along with a step-by-step guide which must be run through each period to produce the data which is provided to ORR and uploaded onto the NR NPT's SharePoint site.

The data sources and processes used for the calculation of the On Time measure at national, regional and operator level are summarised in the Process Map shown in Figure 4.

Figure 4 – Process for producing the On Time measure at national, regional and operator levels.



The On Time measure is based solely on data within TRUST which is fed into PSS. Data is extracted from the 'On Time' Universe in PSS using a Business Objects query which filters only on train services which are relevant to the measure, so those trains:

- with an applicable trains flag, i.e. scheduled to run in passenger service (i.e. revenue earning) as at 22:00 the day before operation; and
- which are run by a government funded operator or by one of four Open Access Operators as described in Section 2.2.1.

The query also filters only on Applicable Stops as described in Section 3.1.

Using this query ensures that data is extracted in a consistent manner each period, including historical data for refresh purposes. The data extracted from this query is then uploaded into a spreadsheet from which the National, Regional and Train Operator On Time Measures are calculated. Some simplistic manual processing is undertaken at this stage, primarily copying and pasting data between files, and then copying formulae down in certain worksheets to update for latest periods. NR has advised the Reporter Team that plans are in place to automate these processes in the coming months.

The latest data on the On Time measure is stored on Network Rail's Performance SharePoint site which can be accessed by industry performance practitioners. This SharePoint site also provides Power-BI dashboards BI which enables users to visualise the data. The data refresh policy is identical to that for the Cancellations Measure, as described in Section 2.3.

The same governance structures as outlined for the Cancellations measure are in place for the punctuality measures, through use of the Taskmaster system, with step-by-step guidance provided, and a team of 4 people who are fully versed in the production schedule.

The Performance Reporting Manual describes quality assurance activities that should be undertaken on the Punctuality Measures. As well as the activities covered in the Cancellations section, NR has recently introduced a check to ensure that the stations being used to calculate the On Time measure (and associated

measures) are accurate. For example, this will check whether any junctions, or other non-station timing locations, are being included in the measure. This check is described in a Process Overview and Work Instruction document which was provided to the Reporter Team ("**Reported Station Assurance Check (Simple**)", version 1.0, dated 12th July 2022).

An example output from this quality assurance check was provided to the Reporter Team for Period 9 of 2023/24 ("**Incorrect Station List P09**"), the first time that this check had been carried out. This identified a total of just 36 station stop records in the period which did not relate to a station (out of c5.6 million total recorded station stops in the period), so no material impact on reported figures.

NR did note that this first check also highlighted a separate issue; a station in Scotland which was not assigned to Scotland Region in the station master list. Therefore, this check also now enables any issues with this station master list to be flagged up, so providing confidence that the Measure is not including any erroneous records which may materially affect the result.

As with the Cancellations measures, we observe that NR appears to have adopted a well-structured process, supported by the Taskmaster system, which provides clarity on progress of core activities required to produce the measures, alongside clear guidance documentation. Checks are put in place to assure the data that is being reported. In our view, this supports the reliability, quality and consistency of the production of the reported measures.

3.4 Data Accuracy for On Time Measure

To review the accuracy of the On Time measure, we have undertaken three core tests:

- Validate the filtering and extraction process used by NR to ensure that all valid records are being captured, and no erroneous records are feeding into the measures.
- Validate the current On Time process to convert PSS data into published On Time figures.
- Verify that assurance has been undertaken following a change in system architecture within PSS in 2022.

3.4.1 Validation of Extraction of PSS Data

In the same way as outlined with the Cancellations measure in Section 2.4.1, we have replicated the filtering process for the On Time measure based on the same sample of 5 days of unfiltered PSS data.

For each of the 5 days, we have identified records that would be excluded based on the filter and confirmed this is appropriate. From our independent filtering of the data, we have compared the number of records with that used for the official reporting for those five days to confirm that these match directly.

This is summarised in Table 5 below, along with a breakdown of the types of record excluded based on using the defined filtering process in NR's documentation. On this basis, we are confident that the filters are being applied as intended and no records are being erroneously omitted. We note that this conclusion is based on the assumption that the data held within TRUST is correct (i.e. we have not examined the quality of the TRUST data since it is out of scope for this study).

Table 5 – Replicated sample of daily figures for On Time measures.

	16th June 2023	22nd November 2023	5th December 2023	16th December 2023	21st December 2023
Total Records (unfiltered dataset)	794,790	805,182	880,263	745,718	813,989
Excluded Records:					
Non passenger Operator	66,739	70,550	70,920	39,124	62,992
Non applicable train	2,700	2,905	4,480	4,441	8,253
Non PfPI train	44,995	47,355	49,513	33,449	45,379
Train running on excluded TOPS code	130	359	74	46	71
Pass or Depart record	407,664	410,270	451,167	401,067	418,091
No GBTT time for stop	4,065	3,892	3,659	3,465	3,762
No actual time for stop	22,448	20,259	87,913	35,912	38,232
Not a recorded stop (fail to stop)	436	604	430	413	659
Total Excluded Records	549,177	556,194	668,156	517,917	577,439
Comparison of recorded stops: Arup calculations vs NR official data					
Recorded Stops (Arup)	245,613	248,988	212,107	227,801	236,550
Recorded Stops (NR)	245,613	248,988	212,107	227,801	236,550
Difference	0	0	0	0	0

The figures in the "Not a recorded stop (fail to stop)" row are those records where a valid train was operating and a time was recorded at the station stop, but a Fail to Stop event was assigned to that station stop. These are the records which were included in the On Time measure prior to the March 2022 update, but which are now correctly excluded. Excluding these reduces the number of station stops used in the measure by c.0.2%.

As shown in Table 5, for all dates were able to directly match the number of records being used to those in the published statistics. We were also able to directly match the calculated On Time, Time to 3 and Time to 15 measures for these 5 days, so providing confidence in the accuracy of the reported figures.

3.4.2 Validation of Calculation of On Time Measure from PSS Data

Process notes were provided to document the current procedure for inputting and calculating PSS outputs to reported numbers for the On Time metrics at National, Region, Route and Train Operator level.

Our tests covered the following areas:

- Validation that data extracted from PSS ('Punctuality – Current CP.xlsx') was correctly copied into appropriate locations in the 'Punctuality – Official.xlsx' workbook.
- Formulae within the 'Punctuality – Official.xlsx' workbook were correctly dragged down in accordance with instructions.
- Confirmation that calculated Moving Annual Average (MAA) values within the 'Punctuality – Official.xlsx' were replicable.
- Consistency checks for the different types of grouping for reporting within the 'Punctuality – Official.xlsx' workbook.
- Testing metadata cells are updated workbook.

All tests that were undertaken were completed successfully, demonstrating that the production of the values is correct. The following observations were made which are suggested as recommendations to improve the process:

- Redundant formulae cells were identified on the 'Operator MAAs' sheet of the 'Punctuality – Official.xlsx' workbook.
- Broken references were identified on the 'ChartData' sheet of the 'Punctuality – Official.xlsx' workbook.
- In line with best practice, cells which contain formulae should be made identifiable from copied values, for example through the use of coloured cells.

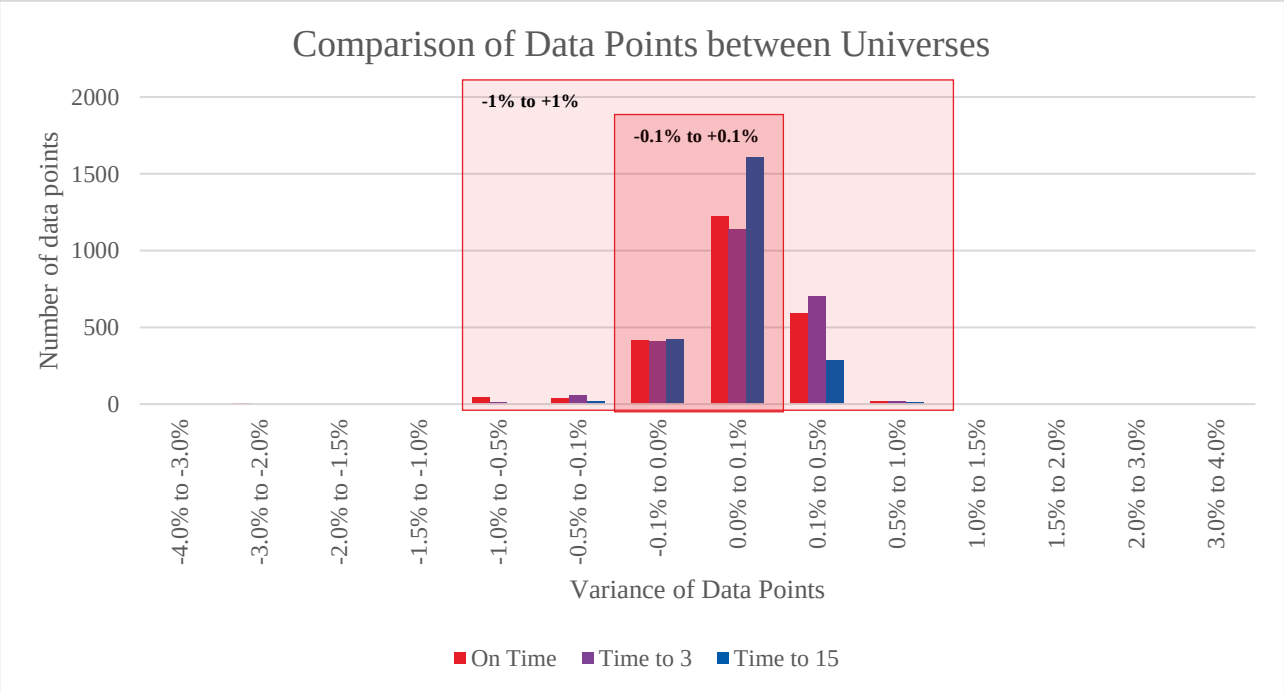
3.4.3 Validation of Universe Changes

As outlined in Section 3.2.2, NR updated the approach for calculating the On Time Measure, to better reflect Fail to Stop records. This included the creation of a new Data Universe within PSS ("On Time Universe") (OTU) which also provided efficiencies in data extraction.

As part of this update, NR undertook an internal validation exercise and compared how the On Time measure would change as a result of this update to the new approach . To support our review, NR provided their comparisons (in file '*DOLARS v OTU (March 2022) v OTU (Dec 2023).xlsx'*). The comparison period covered 2014/15 Period 1 to 2021/22 Period 12, with each data point based on individual Operators, giving a total of 2,350 points. 99% of the revised On Time measure figures were within 1% of the original reported figure. 70% of On Time neasure showed a variation within +/- 0.1%. The the maximum underreported change seen for an individual Operator (Elizabeth Line) in a period was 2.4%, and maximum overreported change for an individual Operator was 4.0% for the Caledonian Sleeper, noting that Caledonian Sleeper operates a smaller number of services.

Reviewing each of the data points (based on individual Operators in each period, so 2,350 data points), Figure 5, below plots the histogram distribution of data points by percent difference bands.

Figure 5 – Comparison of data points between old and new approaches.



3.5 Confidence Rating for On Time Measure

Our audit has identified no material concerns with the reliability of the reported On Time measure (and associated Time to 3 and Time to 15 measures). The use of automated queries to extract data ensures a consistent approach, while the process is well documented with further resilience provided by the size of the team available to run the process.

Reflecting on our review, the On Time Measures are awarded **'A'** for overall reliability. The same process is applied for National, Regional and Operator On Time so this score reflects all three measures.

Reflecting on our review of the accuracy and processes undertaken, the On Time Measures are awarded a '**1***' for overall reliability. We have been able to independently replicate the values reported to within 0.1% of published figures.

It should be noted that these scores cannot be directly compared with the scores awarded in the 2017 review. The current review focuses solely on the conversion of TRUST data into the reported measures, and so the scores reflect our views of the reliability and accuracy of this process. The reliability and accuracy scores from the 2017 review also heavily drew upon issues observed with the capture of data within TRUST, in particular in relation to berthing offsets. TRUST data (including berthing offsets) is outside the remit of this review. If the scope of the 2017 review was similar to the current review, then the scoring is likely to have been more consistent with this review.

4. Recommendations from Previous Independent Reporter Review

Table 6 below summarises the relevant recommendations from the 2017 audit of these measures and our view on progress against these. Note, recommendations related to capture of data within TRUST are not covered below, since outside the scope of this project. As such only 2 of the 15 recommendations from that review have been covered in this project.

Table 6 – Status of relevant recommendations from the 2017 audit

Reference	Recommendation from the 2017 audit	Status
2017NPM01	Enhance the documentation for the calculation of all metrics, produce a Record of Assumptions, and apply best practice guidelines to the spreadsheets. The documentation should include the processes used to generate source data in BO-PSS for relevant metrics.	Evidence provided of clear documentation for all Measures, including process for generating source data. We have noted there remains scope to apply best practise guidelines to spreadsheets.
2017NPM07	Ensure that there is a defined list of reporting stations for On Time and that there is a formal process for the addition or removal of stations	The recommendation was originally intended to reflect a potential impact on the measures as the number of reporting locations increased. With the proportion of stations now reported against at c.97% (compared with c.80% at the time of the 2017 review), this issue is no longer relevant and so this recommendation is closed.

5. Conclusions and Recommendations

The following Table 7 outlines a summary of the confidence grades for each of the measures considered in this review based upon the analysis outlined in this report.

Table 7 – Summary of confidence grades

Measure	Current Confidence Grade	Commentary
National and TOC Cancellations Measure	A1*	Our review of NR documentation, data and processes has not identified any material concerns with the reliability of the reported National and Operator Cancellations measure. We have also been able to replicate the reported figures to within 0.1% giving an accuracy score of 1*.
On Time Measure (and Associated Time to 3 and Time to 15 Measures) for National, Regional and Train Operator	A1*	Our review of NR documentation, data and processes has not identified any material concerns with the reliability of the reported National and Operator Cancellations measure. We have also been able to replicate the reported figures to within 0.1% giving an accuracy score of 1*.

As the Regional Cancellations measure is not yet formally produced and as part of the review we have therefore not been able to conduct all tests, the following Table 8 presents provisional scores on the assumption that the process for reporting the Regional Cancellations builds on existing process as described in our review. We would also expect NR complete its planned update for the Definition documentation.

Table 8 – Provisional confidence grades

Measure	Current Confidence Grade	Commentary
Regional Cancellations Measure	A1* *Provisional*	We do not expect there to likely be any material concerns with the reliability of the Regional Cancellations measure on the assumption that the process builds on the NR documentation, and processes reviewed for existing measures. We have also been able to replicate the reported figures to within 0.1% giving a provisional accuracy score of 1*. As the Regional Cancellations measure is not currently reported, it has not been possible to conduct Test 2 for this measure.

Upon completion of this review at this time we do not have any specific recommendations to be made under the scope of the mandate, however the following Table 9 outlines suggestions for potential improvements to be considered.

Table 9 – Suggestions for potential improvements

Reference	Advisory Suggestion	Benefit	Owner
2024.35532 .Adv01	The formulae used to calculate moving annual average figures in the 'TOC Data' sheet are based on direct cell references to previous rows. There is a risk that if the order that data is extracted from PSS changed in one period or if the number of Operators for which data is extracted changes, this could result in incorrect calculations. We would propose a generic consistent formula is applied based on the relevant TOC and the Period indices (to identify the relevant periods for the calculation).	Reduced risk of errors where changes may occur to extracted data from PSS or changes to the number of operators.	Network Rail
2024.35532 .Adv02	We suggest it would be useful to build some internal validation into the workbook to support validation of the metric, for example a check for duplicate values (e.g. if data for a particular period and TOC is manually pasted into this worksheet twice).	Improved assurance process to check for any possible data errors.	Network Rail

Reference	Advisory Suggestion	Benefit	Owner
2024.35532 .Adv03	There were some inconsistencies identified in the formulae in the 'TOC Data' sheet. It is advised that a visual reference (e.g. different coloured columns) be applied to highlight this difference. For example, inconsistent formulae were observed in Column I of the "TOC Data" sheet between rows 2281 and 2283 when the Lumo Operator was first introduced (this links to 2024.35532.Adv01 above, where it is suggested that a more generic consistent formula is applied).	Reduced risk of errors being introduced to the spreadsheet	Network Rail
2024.35532 .Adv04	Include within the Definitions of Railway Performance Metrics for Regional Cancellations any finalised methods relating to the handling of services indicated as part or fully cancelled but exclude flags indicating which stops have been cancelled.	Provides greater clarification on these specific cases.	Network Rail

A.1 Statement of Work (SoW) #37788

Independent Reporter Framework

Statement of Works

COMMISSION INFORMATION	
Project Name:	Train Performance Measures
Bravo Contract Number:	#37788
Network Rail Contact:	[redacted]
Network Rail Department:	Planning & regulation
Date Raised:	13th July 2023
SoW Number:	[insert SoW number – C&P only]
Network Rail PO Number:	[insert NR PO# when available]
Commission Value:	[insert the SoW value after this has been agreed with the supplier]
Supplier Name:	[insert the name of the selected supplier after appointment]
Main Supplier Contact:	[name and email address of the main supplier contact]

This Statement of Work (SoW) is the contractual vehicle for defining, authorising and commissioning a piece of work to be undertaken under the Independent Reporter Framework. The SOW has six sections:

1. *Commission Information*
2. *Commission Overview*
3. *Scope of Services and Deliverables*
4. *Knowledge Transfer*
5. *Commercial Details*
6. *Performance Measurement*

This SoW is entered into under and in accordance with the terms of the Independent Reporter Framework dated 1 February 2020 between Network Rail, the Office of Rail and Road, and the Supplier and includes and incorporates any special Terms and Conditions and any other amendments captured in this SoW.

Any dispute surrounding this SoW will be resolved in accordance with the Terms and Conditions outlined in the Framework Agreement.

Ownership and use of any Intellectual Property Rights shall be in accordance with the Framework Agreement Terms and Conditions.

Change control procedures are to be applied as set out in the Terms and Conditions of the Framework Agreement.

COMMISSION OVERVIEW	
2.1 Background	On Time (and associated measures – Time to 3, Time to 15) and the cancellations measures are currently used by and published on both Office of Rail and Road (ORR), and Network Rail's websites as part of the monitoring of passenger train performance. In the latest Periodic Review (PR23) policy framework, these measures will be used as the passenger train performance success measures for Control Period 7 (April 2024 to March 2029). These performance measures (together with a basket of other measures) were first introduced in 2017 by the rail industry to improve rail performance by providing the right incentives for Network Rail and train operating companies, and also to increase customer satisfaction by promoting transparency in terms of the impact of train performance on passengers. On Time, Time to 15 and Cancellations measures were last reviewed by an Independent Reporter in July 2017 for system reliability and data accuracy as part of a wider review of the new performance metrics. On Time and Time to 15 scored C1 and Cancellations scored B2 (see appendix 2 in the report for the grading system). Following implementation of a new methodology in March 2022 for On Time and associated measures, it is deemed an appropriate time to re-assess the accuracy and reliability of these measures. Definitions of the measures for the review: On Time measures the percentage of recorded station stops* arrived at early or less than one minute after the scheduled time. * A recorded station stop is defined as a location with both a planned timetable time and an actual recorded time where a train has stopped. The associated measures Time to 3 and Time to 15 measure the percentage of recorded station stops arrived at early or less than three and 15 minutes respectively after the scheduled time. Cancellations measures the percentage of planned trains which either did not run their full planned journey or did not call at all their planned station stops. This measure is a score which weights full cancellations as one and part cancellations as half. A lower cancellations score indicates better reliability. This industry measure is an indicator of disruption against the timetable operating on the day. The timetable is finalised at 22:00 the previous evening, and trains removed from the timetable before then will not be included.
2.2 Business Objectives and Priorities	The objective of the Independent Reporter review is to measure the system reliability and data accuracy of On Time (and associated measures – Time to 3, Time to 15) and cancellations measures. As part of the important role in monitoring train performance and being success measures in CP7, it is critical that Network Rail, ORR and rail industry stakeholders have assurance of the quality of the data and robustness of these measures. The review should: - build on the review of system reliability and data accuracy of these measures undertaken in July 2017; - consider the suitability of the improved methodology for calculating On Time and associated measures which was implemented by Network Rail in March 2022;

	- provide updated assessments of system reliability and data accuracy for On Time (and We do not expect the review to include: - assessment of the quality of data in TRUST; - any focus on pre-plan cancellations or p-coding.

3 .0 SCOPE OF SERVICE AND DELIVERABLES	
3.1 Key requirements	The Reporter will be required to assess the system reliability and data accuracy of the On Time (and associated measures – Time to 3, Time to 15) and the Cancellations measure. The Reporter will be required to review, comment and make recommendations on the: - governance for collecting the data for the On Time (and associated measures – Time to 3, Time to 15) and Cancellations measures in TRUST and the improved methodology (in March 2022) for transforming it into outputs provided across the rail industry; - reliability, quality, consistency, completeness and accuracy of reported data for On Time (and associated measures - Time to 3, Time to 15) and Cancellation measures at a national, Network Rail region and train operator level; - processes to produce, quality assure and provide consistent period end data to customers/stakeholders including ORR. The reporter will also be required to; - present a confidence grading; both an alpha (system reliability) and numeric (data accuracy) grading based on the most up to date dataset available during the commission (see section 1.6 below); and - make recommendations for improvements Network Rail would need to make to achieve higher gradings (if appropriate). The On Time (and associated measures) and Cancellation measures are produced centrally by the Network Performance Team (NPT) at Network Rail. Therefore, we expect this commission to require working primarily with NPT. To pre-empt any potential recommendations and minimise the risk of duplicating work, the Reporter should work with Network Rail and ORR to understand any known issues or existing work streams that could impact on potential gradings. The Reporter should familiarise themselves with the background to these measures. Additionally, the Reporter should review previous Independent Reporter

	reviews that have covered accuracy of data out of TRUST, including: - Right Time Performance review (2013) - Review of performance measures (2013) - Review of new performance metrics (2017) - Review of CRM-P and FDM-R (2021)
3.2 Key skills	Bidders will need to demonstrate how they meet the key following skills and experience: • have access to suitable tools and software in order to provide the detailed analysis • capable of producing a reliable and efficient method for analysis and assessment • the ability to work collaboratively with key stakeholders at all levels
3.3 Key deliverables	**- a confidence grading on both the system reliability and data accuracy for each of the measures in line with the grading system below;** **- regular progress updates, including early identification of any potentially significant issues identified as part of the work;** **- a presentation of draft findings to be discussed at a meeting with Network Rail and ORR;** **- a draft report (for comment by ORR and Network Rail) covering the issues set out in the scope section above by 5th February 2024; and** **- a final report that addresses comments provided by ORR and Network Rail on the draft report by 23rd February 2024.** **System reliability grading system** **Band** / **Description** A: Sound textual records, procedures, investigations or analysis properly documented and recognised as the best method of assessment. B: As A but with minor shortcomings. Examples include old assessment, some missing documentation, some reliance on unconfirmed reports, some use of extrapolation. C: Extrapolation from limited sample for which Grade A or B data is available. D: Unconfirmed verbal reports, cursory inspections or analysis • **System reliability is a measure of the overall reliability, quality, robustness and integrity of the system that produces the data.** • **Some examples of the potential shortcomings include old assessment,**

missing documentation, insufficient internal verification and undocumented reliance on third-party data.

Accuracy grading system

Band	Description
1*	within 0.1%
1	Data used to calculate the measure is accurate to within 1%
2	Data used to calculate the measure is accurate to within 5%
3	Data used to calculate the measure is accurate to within 10%
4	Data used to calculate the measure is accurate to within 25%
5	Data used to calculate the measure is accurate to within 50%
6	Data used to calculate the measure is accurate to more than 50%
X	Data accuracy cannot be measured

- **Accuracy is a measure of the closeness of the data used in the system to**

the true values.

- **Accuracy is defined at the 95% confidence level - i.e. the true value of 95%**

of the data points will be in the accuracy bands defined above.

3.4 Proposed approach	*[Insert at contract award stage]* [Demonstrate and detail the proposed approach for the project, covering all areas of the projects scope and clearly state the requirement(s)]
3.5 Schedule & timings	Purposed Contract Start Date: 15th December 2023 *These are indicative dates and will be agreed once the contract has been awarded and the PO has been approved. [Insert details pertaining to the commission's intended start and end date, as well as a commission schedule e.g., a Gantt chart with tasks and attributive start/end dates]
3.6 Relationship applicable for performing the duties under this statement of works contract	Data Controller and Data Processor. The only processing that the Supplier is authorised to do is listed as in Appendix 1 and may not be determined by the Supplier.

4.0 KNOWLEDGE TRANSFER	
4.1 Knowledge Transfer	*[Insert at contract award stage]* [Explain and detail how knowledge transfer is to be enabled throughout the commission and how the final output will be delivered and presented to Network Rail and ORR.]

5.0 RESOURCE & COMMERCIAL DETAILS	
5.1 Supplier Resource	*[Insert at contract award stage]* [Key personnel which will be engaged in the commission, along with their responsibilities. Details should include sub-contractors, if sub-contractors are being utilised for the delivery of this contract commission] In the event of "key personnel" becoming unavailable the supplier agrees to provide a replacement of equal standard and status within 48 hours of notice.
5.2 Pricing Schedule	**This contract is based on a FIXED PRICE lump sum.** *[Insert price schedule and cost breakdown at contract award stage]*
5.3 Payment Milestones	This contract is being let on a FIXED PRICE contract, payable on completion of key milestones and detailed at contract award stage.
5.4 Place of work	Network Rail's NPT team are mainly based at The Quadrant, Elder Gate, Milton Keynes MK9 1EN. For the purpose of this contract, this will be deemed the usual place of work. It is expected however that the successful Supplier will deliver most of the work from the Supplier's own premises. Travel for various meetings, workshops and engagements is likely to include, but not limited to:- • Network Rail's offices in Milton Keynes • Regions as required Business Travel Expenses to UK locations **other than** the Reporter's normal place of work may be claimed, subject to prior agreement and in accordance with Network Rail 's Business Travel and Expenses Policy. The Supplier shall endeavour to minimise travel and expense costs throughout the duration of the contact.

	Currently Network Rail is utilising remote working facilitated by video-conferencing platforms such as Microsoft Teams. Therefore, it is anticipated that the Supplier will be able to use this tool.
5.5 Expenses	For the purpose of this contract, business travel expenses to Network Rail, TOC or FOC offices [if this becomes necessary] may be claimed in accordance with Network Rail's Business Travel and Expenses policy.
5.6 Contract Variations	Variations to this Statement of Work contract may be permitted in accordance with Clause 88 of the Utilities Contract Regulations (modification of contracts during their term). All variations to this Statement of Work contract must be agreed in writing under a restated statement of works document, duly signed by all parties

6.0 INVOICING	
6.1 Invoice Details	Network Rail operates a strict "NO PO – NO PAYMENT" policy. Invoices are to be raised on completion of the contract or in accordance with the milestone payments [where applicable] set out in this SOW. Invoices should contain the following information as a minimum: • Purchase Order number • SOW number as detailed in Section 1.0 • Project Title and description Business expenses should be invoiced as a separate line and supported with receipts, as described in terms and conditions of the framework agreement and the Network Rail Business Expenses Policy. Please be aware that failure to provide the information above may potentially cause a delay in processing the invoice. Our preference wherever possible, is for invoices to be submitted via EDI. Alternatively, invoices may be submitted By email - invoices@networkrail.co.uk By post – Network Rail Accounts Payable, PO Box 4145, Manchester M60 7WZ

This Statement of Work will be executed as per the Terms and Conditions agreed in the Independent Reporter Services Framework Agreement.

[supplier name to be completed at contract award]

Signed:……..

Name (CAPS):……

Position:……

Date:……

NETWORK RAIL

Signed:……..

Name (CAPS):……

Position:……

Date:……

[This SOW does not require further contract signatures from the ORR]

ANNEX 1 – Protection of Personal Data

Where Data Controller and Data Processor applies

The Supplier shall only process personal data as detailed below:

Description	Details
Data Protection Officers	**Network Rail**: [redacted], The Quadrant, Elder Gate, Milton Keynes, Buckinghamshire, MK9 1EN **Supplier: inserted at contract award stage**
Subject matter of the processing	The processing is needed to ensure that the Processor can effectively deliver the services under the framework contract.
Duration of the processing	The duration of processing refers to the duration of the contract, as specified in the call-off contract
Nature and purposes of the processing	The nature of the processing means any operation such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction of data (whether or not by automated means). The purpose might include (but not limited to): statutory obligation, arranging Stakeholder meetings, data research and analysis and compliance with Network Rail's Business Travel and Expenses policy.
Type of Personal Data being Processed	This may include (but is not limited to): name, address, job title, location, email address, telephone number, images, cost center number biometric data.
Categories of Data Subject	Examples include (but is not limited to): staff (including sub-contractors, volunteers, agents), customers/ clients, suppliers, students, apprentices, members of the public, users of a particular website.
Plan for return and destruction of the data once the processing is complete UNLESS requirement under union or member state law to preserve that type of data	On completion of the processing (interpreted as being contract expiry) the supplier shall cease to use the personal data and shall arrange for it's prompt and safe return to Network Rail, or destruction if instructed by Network Rail, of all Personal Data.

A.2 Confidence rating system

A.2.1 Reliability grading system.

System reliability band	Description
A	Sound textual records, procedures, investigations or analysis properly documented and recognised as the best method of assessment.
B	As A but with minor shortcomings. Examples include old assessment, some missing documentation, some reliance on unconfirmed reports, some use of extrapolation.
C	Extrapolation from limited sample for which Grade A or B data is available.
D	Unconfirmed verbal reports, cursory inspections or analysis.

Notes:

1. System reliability is a measure of the overall reliability, quality, robustness and integrity of the system that produces the data.
2. Some examples of the potential shortcomings include old assessment, missing documentation, insufficient internal verification and undocumented reliance on third-party data.

A.2.2 Accuracy grading system.

Accuracy Band	Description
1*	Data used to calculate the measure is accurate to within 0.1%
1	Data used to calculate the measure is accurate to within 1%
2	Data used to calculate the measure is accurate to within 5%
3	Data used to calculate the measure is accurate to within 10%
4	Data used to calculate the measure is accurate to within 25%
5	Data used to calculate the measure is accurate to within 50%
6	Data used to calculate the measure is inaccurate by more than 50%
X	Data accuracy cannot be measured

Notes:

3. Accuracy is a measure of the closeness of the data used in the system to the true values.
4. Accuracy is defined at the 95% confidence level - i.e. the true value of 95% of the data points will be in the accuracy bands defined above.

A.3 List of reviewed documents

Document	Date Received
20220323_Letter_ORR response to NR_2022_23 regional comparison scorecard changes_FINAL	2023-12-15
On time and Calendar Universes How To	2023-12-15
Cancellations.docx	2023-12-15
NR-NPRT-WI-004 (Severe Disruption and Cancellations).docx	2023-12-15
NR-NPRT-WI-005 (Punctuality - Official) - Provisional.docx	2023-12-15
NR-NPRT-WI-010 (Punctuality - Official) - Confirmed.docx	2023-12-15
Punctuality.docx	2023-12-15
29 - Metrics Reporting Manual_V1	2023-12-15
Definitions of Railway Performance Metrics_v3.03c	2023-12-15
20220323_Letter_ORR response to NR_2022_23 regional comparison scorecard changes_FINAL.pdf	2023-12-15
Definitions of Railway Performance Metrics_v3.03c.docx	2023-12-15
On time and Calendar Universes How To.docx	2023-12-15
On Time Methodology Note SD.pptx	2023-12-15
Definitions of Railway Performance Metrics_v3.03b.docx	2023-12-17
File Sync.docx	2023-12-17
RE_ ORR Independent Reporter Review - Performance Metrics _Filed 17 Jan 2024 16_38_.msg	2023-12-17
Regional Cancellations Definition.docx	2023-12-17
Regional Cancellations in BOPSS.docx	2023-12-17
Reported Station Assurance Check (Simple).docx	2023-12-17
TaskMaster&Worklist Process.docx	2023-12-17