Smartphone Cities

Technical annex

Research Document

Publication date: 16 December 2016

Section 1

Our metrics

1.1 Introduction

In this section we aim to explain the metrics that we have measured. We also describe briefly our benchmarking tool as well as important points to bear in mind when reviewing results obtained from our testing. We also explain the reasons for exclusion of some of the results from our gathered measurement.

1.2 Measured parameters

We measured metrics on 6 different tests types which will believe gives a good measure of network performance on consumer smartphone activities as listed below.

- HTTP Download speed and success rate
- HTTP Upload and success rate
- Ping response time
- Web browsing speed and success rate
- YouTube;
    - Time to first picture
    - Freezing time
    - Jerkiness
    - Playback Resolution
    - Video Quality
- Voice Call Setup Time, Success rate and Quality

It is considered that these metrics capture important aspects of mobile broadband network performance. For this reason, the results are intended to be used as part of the information required by the consumer when making choices concerning mobile broadband provision. The following sub-sections explain each of the metrics in more detail:

HTTP Download Speed and Success Rate

HTTP (hypertext transfer protocol) is the method commonly used to transfer information over the internet; for example, in delivering web pages or a video stream. Download speed indicates the rate at which a connection is able to transfer data from the internet to the consumer.

A connection with a higher download speed would take less time to transfer the same data than a connection with a lower download speed. For example, at a constant speed of 20Mbit/s (20 million bits per second), the theoretical time taken to download a 10MB

(83,886,080 bits of data) file would be just over four seconds, while on a constant speed of 10Mbit/s, it would be just over eight seconds.

The rate of HTTP download can vary according to the location and time of day, even during a short session of use, for a multitude of reasons. Some of these effects may be related to the mobile network that is providing the connection to the internet such as contention on the cell the device is registered to, or the operator’s use of traffic management systems, the handset being used to access the service or the SIM card tariff used by the consumer.

It can also be limited by factors outside the control of the mobile network operator. For example, if multiple users are attempting to access content from the same server at the same time, and that server lacks the capability to serve them all at the same time, the download rate could be limited.

HTTP is used to deliver many types of content, including web pages, audio, video, and images, as well as for downloading applications to a consumer’s device. While HTTP is used for delivering various types of content, this content may be treated in different ways by the content providers and by the networks that transmit the information.

The download success rate provides an indication of the number of times a request to download a file was made through the test equipment versus the number of instances that the request to download was successful. This is expressed in the results as both a percentage; e.g. if 20 requests for a download were made and 19 were successful, the success rate is 95%.

Response Time

Response time (referred to technically as round-trip time or latency) indicates the delay between a request for information and the response. A connection with low latency will “feel” more responsive and certain applications perform far better with lower latency.

Latency was measured by sending a series of ICMP (internet control message protocol) “ping” tests. Latency refers to the responsiveness of a network and is measured as the time between sending a signal and receiving a response from the targeted system. An example to show the effect of latency is demonstrated through live satellite television news broadcasts, where a delay is sometimes seen between a presenter asking questions in a studio in the UK and the response from the reporter in a distant location.

Low latency is important for applications that require information to be delivered with as little delay as possible. In particular, low latency is most important when using services such as video calling, VoIP (voice over internet protocol) and online gaming.

HTTP Upload speed and success rate

As with download speed, upload speed indicates the rate at which a connection is able to transfer data from one device to another, although with upload speed this represents the rate at which data can be transferred from the handset to the remote Data Server.

The Upload speed success rate is defined in the similar manner as the HTTP Download success rate, however in this case referencing the success of the upload requests.

Web page loading speed and success rate

Web browsing speed indicates the amount of time it takes to completely load a given page hosted on a website. We chose to use three of the most popular consumer webpages for testing as well as a standard static HTML reference webpage:

- The BBC, Amazon and YouTube homepage are examples of a commonly used webpage with dynamic content i.e. the actual content of the page changes on a regular basis. As a request from a smartphone to download these webpages often leads to a redirection to the mobile version of these sites, we have carried out our testing on the on the mobile version of these websites in order for our results to be more representative.

- A standard HTML reference web page. The page used for testing is based on an ETSI (European Telecommunications Standards Institute) "mKepler" standard reference page, designed for smartphones to represent a typical static (i.e. unchanging) HTML web page.

For testing, we ensured no URL redirection has taken place as part of the page loading sequence for both websites. Each webpage was loaded completely, with no cache held on the devices to test network performance. Consumer devices typically cache parts of webpages to contribute towards a faster download experience.

Each approach has its own advantages and disadvantages. Using a standardised reference web page hosted on a dedicated server means that the conditions for downloading this page remain stable for each test, however as this page type is not typically accessed by consumers it could be viewed as unrepresentative.

Loading a web page with dynamic content (e.g. the BBC Homepage) does provide a more typical example of how a consumer would normally use a mobile device, however this also introduces another variable into the equation as the page content size is not guaranteed. There is also a possibility that unusually heavy use of the website (such as during a major breaking news story) may also affect performance. Regardless of the web page chosen, there was the possibility that the mobile operators would have optimised the data transferred to the handset as part of their normal operations.

In the same manner as the HTTP Download and HTTP Upload success rate, webpage loading success rate is ratio of the number of successful requests to the total number of requests made.

YouTube (Video streaming)

The video clip used for testing was one of the most popular videos played in the UK at the beginning of testing in the Summer of 2016 - the film trailer for "Absolutely Fabulous". It is important to note that the size and quality of a video file streamed is determined by the content provider depending on a number of factors including the device capabilities and the network capacity at the time that the content is being delivered. The network may choose to re-encode this content before providing it to the end-user. This is in order to create a more responsive experience by minimising the volume of data transferred and use of its network capacity. The decision by the operator to perform this will be reflected in the metrics recorded.

<u>Time to first picture</u>

This is defined as the time between making a "play" request to the content provider and the time the first image of the requested video clip is displayed on the device.

Freezing time

This is defined as the cumulative time that the requested video clip freezes during playback. As this is a cumulative statistic, a result of 1.5 seconds may be derived from one pause of that length, ten pauses of 0.15 seconds or three pauses of 0.5 seconds etc. Minor freezes that would not be noticed by a consumer are ignored by our measurement system.

Jerkiness

Jerkiness is a perceptual value that measures the loss of information from one frame to the next due to a freezing period or a low frame rate. The value reported is derived from the following measures:

- Freezing;
- Loss of information, estimated by the inter-frame difference;
- Dominating Frame Rate; and
- The amount of time an image remains visible until the image information changes in the next update. In the case of a constant frame rate, the dominating frame rate is equal to the constant frame rate

Playback Resolution

During a video session, there is a negotiation between the mobile device and the content provider to determine the optimum video resolution delivered to the mobile device. The resolution of the first picture to be displayed (for example 360p or 720p) is then recorded, giving an indication of the video resolution during the session.

Video Quality

This metric reports a number between 1.0 (bad) and 5.0 (excellent), derived from the ITU approved J.343.1 ‘RS-T-VModel’ algorithm which analyses the incoming video data for the video client and observes the resultant video at the same time. A video quality score is then calculated based on information taken from the video data, such as the size of a compressed frame, and the visual impression of the decoded and displayed video frame.

Voice Call

We used the following Quality of Service (QOS) parameters to measure the voice performance:

Voice Call Setup Time

This parameter is defined as the time taken between the mobile device initiating a call and the connection to the dialled number being completed, with the call being answered immediately on receipt.

Voice Call Success Rate

All our voice call tests were for a duration of ninety seconds. In the same manner as the other success rate calculations, this metric makes reference to the number of completed voice calls made versus the total number of attempts.

We have defined a “dropped” call as any call that terminates before the ninety second duration, excluding calls that stop due to intervention by our test engineers (i.e. stopping a measurement during a call).

Voice Quality

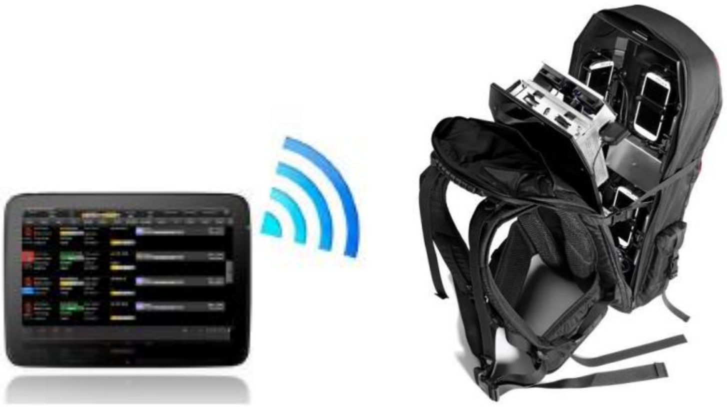
Similar to the Video Quality metric, Voice Quality is reported as a MOS (Mean Opinion Score) between 1.0 (bad) and 5.0 (excellent), using an implementation of the ITU T P.863 POLQA algorithm approved in January 2011. POLQA (Perceptual Objective Listening Quality Assessment) is known as a full-reference model: the quality estimation is based on comparing the transmitted signal with the high quality original reference signal.

We chose to run these tests in “half duplex” mode- measuring the voice quality in both directions in sequence between the measurement equipment and identical handsets controlled by two dedicated Voice Quality Servers situated at Ofcom’s Engineering Hub in Baldock, Hertfordshire.

1.3 Measurement tool

We collected our data using a benchmarking system housing 4 Android handsets. The benchmarking software application runs directly on the handsets we used for testing. A control master tablet device is used to set the test program schedule and sequence for the handsets. The control information from the master tablet to the slave handset is sent over Bluetooth and tests can be observed while they are running.

Figure 1: Benchmarking system with control tablet



The system is usable for both indoor and outdoor measurements (in-building tests, in trains, pedestrian areas, etc.), but portable enough to be put into a vehicle.

1.4 General Rules for using our results

When reviewing our results, it is important to note the following points:

- Indication of better performance: the highest number is not necessarily always the best for all mobile device applications, as the other factors discussed in previous sections can also affect performance.

- Relative difference: even a statistically significant difference may have no user-perceivable impact on the consumer’s QoE.

- Performance can vary between devices.
- The number of 4G subscribers is increasing, which may lead to higher network load and a reduction in performance. Conversely, as 4G networks are further optimised, performance may improve.

1.5 Exclusion of data from our calculations

All the tests were carried out using the test equipment specified earlier to gather data as well as process the results. The outcomes of the tests were broadly classified into

a) Successful – where the test completed successfully.

b) Failed – where the test failed due to various reasons pertaining to the network

c) Aborted – where the tests were unsuccessful in completion due to user intervention or test equipment behaviour.

d) System release – where the individual session was terminated by the control software as part of setting up a particular test.

Both the “Aborted” and “System Release” outcomes were excluded from our results as their inclusion would adversely affect our results.

Section 2

Testing Methodology

2.1 Evolution of test methodology

In this section, we highlight briefly the changes we have made to our methodology in comparison to the previous phase. The rationale behind our changes was to increase the overall measurement sample count per city as well as to maximize reliability of our results.

Choice of cities and measurement area size

Since the last phase, we have increased the measurement area to cover a 10km radius in seven of UK major cities: Southampton, Sheffield, Cardiff, Birmingham, Edinburgh, Belfast and London.

We carried out our testing in two phases for each city, where we measure within a 4km radius in the first phase and then outside the 4km radius but within a 10km radius in the second phase.

The results in the main report reflects measurement in the 10km radius but in the accompanying chart pack, analysis on both 0-4km and 4-10km is presented. The 4km split was to provide commonality with previous testing phases.

In each city, the centre point of the 10km radius was defined, using the major rail station as a guide but adjusted when necessary to capture the maximum possible population within the 10km radius. For example, if the train station was close to the coast, we moved the centre point inland.

The 10km radius for each city was divided into 10 segments to enable an even distribution of measurement results where possible across the city, with each segment covered by a measurement team per day.

Figure 2: Sectored 10km radius area

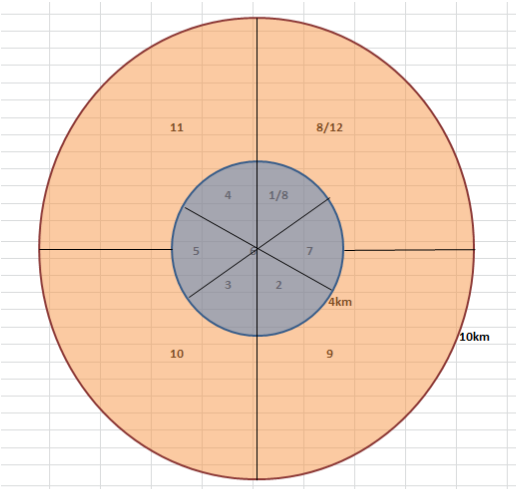
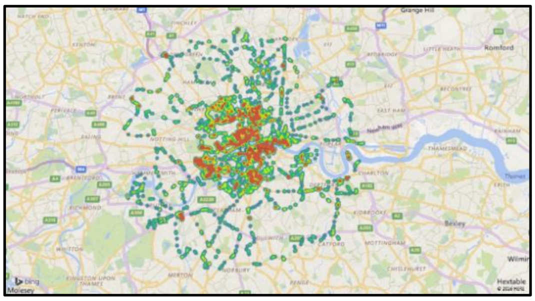


Figure 3: London test area heat map



.

Figure 4: Belfast test area heat map

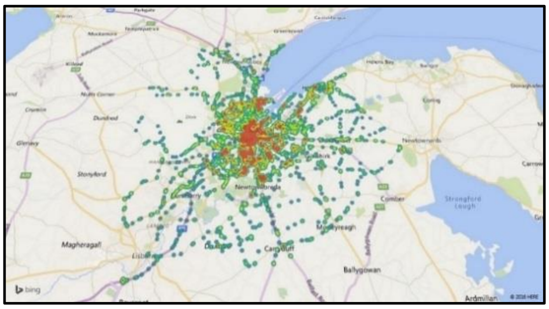


Figure 5: Birmingham test area heat map

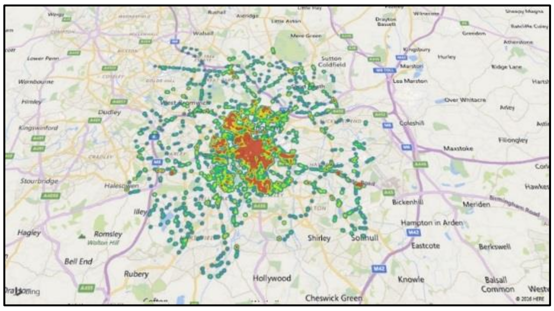


Figure 6: Sheffield test area heat map

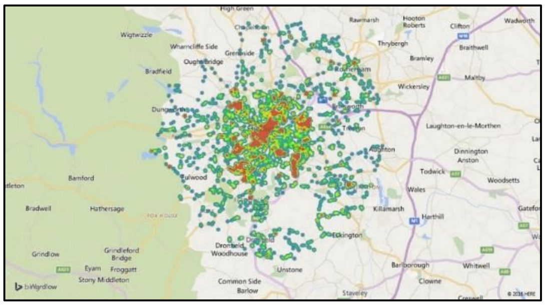
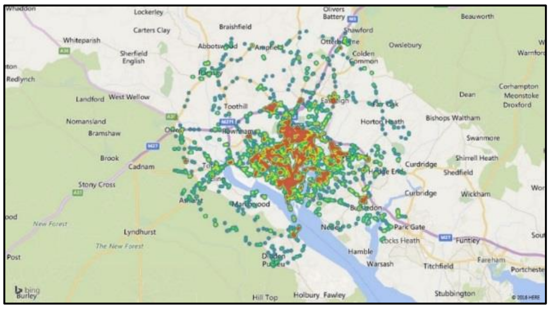


Figure 7: Southampton test area heat map



The breakdown of the measurement test activity within the 4km radius is as follows

- 50% Driving
- 40% Walking with half this time indoors where feasible
- 10% Static

For the 2nd week measurement in the outer radius the measurement activity was split into

- 90% Driving including main arterial roads
- 10% Static

Mobile to Mobile voice call

We have moved in favour of measuring voice quality as well as call success rate from mobile to mobile call setup as opposed to previous phases where we've measured success rate using mobile to fixed lines calls. This we believe is more representative of typical consumer voice call activity. We believe the benefits of accurately depicting user activity outweighs the concern for potential variance that might arise as a result of the receiving mobile end. We've taken steps to minimize any potential for variance in results due to receiving end by ensuring receiving mobile end are positioned in the best possible signal spot for each operator.

Web-browsing

We've expanded the range of our dynamic test webpages to include two more of the most consumer visited websites in the UK aside the BBC which are Amazon and YouTube. The more relevant metric to consumer's web browsing activity we believe is the time taken to download the complete webpage as well as the reliability or the success rate.

Ping

We have changed our approach in measuring latency in the network compared to previous phases. We still maintain the use of pings but a slightly different implementation has been adopted as recommended by the test equipment manufacturer. Our test schedule initially sends a series of ping payloads of larger packet size to ensure the network has settled. We then send the intended test ping payload to measure the latency.

In technical terms, this will ensure the network assigns a high performing bearer in anticipation of more data to be sent through from the handset and therefore enables us to measure the actual latency of the high performing bearer. In our test schedule, our wake-up ping payload is made of 10 pings of packet size 800 bytes and our actual test payload in made up of 25 pings of 32 bytes. We use the average value of the last 5 of the 25 test ping packets to arrive at our measure of latency.

Best bearer measurement

The results produced from the work described here provide indication of overall performance a user with a current handset could expect to see for each operator in the selected cities. The handsets are not forced to measure performance on any specific technology but defaults to the best technology available in the area as determined by the handset. We moved from comparisons between 3G and 4G as we've already demonstrated in previous phases the performance benefits of newer technologies.

Handset choice

We upgraded our measurement test handset to Cat-9 devices capable of utilizing LTE-Advanced or 4G+ technology. These devices are capable of utilizing two or more LTE serving carriers to increase performance. By deploying two or three of these carriers (hence the term carrier aggregation), a network is theoretically able to deliver two or three times the speed of a single carrier. In practise users won't come close to getting the theoretical maximum speed due to constraining factors such as user congestion, backhaul speed, signal coverage and network traffic management.

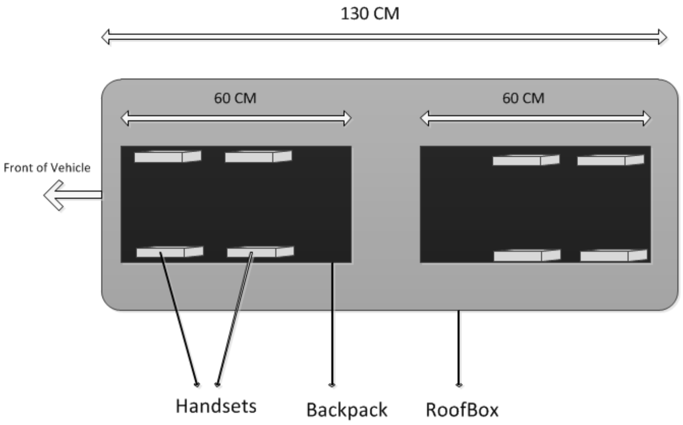
2.2 Scope of Methodology

We have designed our methodology to measure metrics relevant to the consumer experience of using mobile broadband and voice services. It has also been designed to produce a statistically robust dataset that treats each MNO equally. This is to allow us to compare the performance of each MNO's network on a fair and equivalent basis.

As stated previously, we chose a 'best bearer' approach where the handset will choose the best service automatically between 4G, 3G or 2G depending on what is available in that area.

For this phase of testing, we adopted two systems to perform voice and data measurement in parallel for every location. Driving measurements were conducted with the two systems placed in a vehicle mounted roof-box as depicted in 8. Walking measurements were a mixture of indoor (e.g. Shopping Centres) and outdoor measurements along public streets. Static measurements were also a mixture of indoor and outdoor measurements.

Figure 8 Voice and data backpack within roofbox



2.3 Handset Device

The testing for this report was carried out using Samsung Galaxy S6 Edge+ (Model SM-G928F) handsets. This was one of the most popular consumer devices at the time we carried out our tests and was thus reflective of the typical consumer experience of mobile broadband and voice performance.

Handset choice is of course one factor that can affect the network performance experienced by consumers. Early adopters with the latest handsets are likely to experience improved performance as they can benefit from the latest network developments, such as the deployment of carrier aggregation technology from EE and Vodafone and Three's introduction of VoLTE services. Similarly, consumers using devices with older technology may experience poorer performance.

The devices that we used were not MNO-branded devices; i.e. they were not purchased directly from the MNOs. MNO-branded devices generally have firmware pre-installed which is tailored to their network. There is also the possibility of minor customisation to hardware specifications. To allow us to achieve comparable measurements, and to test every network under the same conditions, we did not want to use MNO-branded devices as this would mean that each handset could have been modified in a different way, and would therefore perform in a slightly different way. Using MNO-branded devices would also have precluded us from rotating SIMs across the handsets.

This SIM rotation is important for removing differences between handsets due to manufacturing tolerances. We recognise that there may be small differences in network performance between a branded and unbranded handset, but we consider that the benefits of handset rotation and treating each network equally outweigh any benefits of using branded handsets.

We did not set out to measure the relative performance of different makes or types of device. We wanted to test network performance, therefore as many elements as possible, including the devices used, were set as constants across the testing to maintain comparability across networks.

2.4 Backend data services

To ensure that upload and download transfers from the test handsets to the media server was not the constraining factor in our testing, a high performance server with a 10 Gbit/s IP transit connection was commissioned and tested.

2.5 Data collection tool

The ‘QualiPoc’ application developed by Swissqual was loaded onto each of the handsets and the control tablet to carry out our testing. Swissqual is a provider which specialises in services and systems for measuring, analysing and reporting performance of mobile devices and network services.

The system used for the collection of our results consisted of a backpack containing four handsets and a scanning receiver. The control tablet is used to configure the test setup of the slaves and scanning receiver and to monitor test progress.

At the end of each measurement day, the result data was transferred from the handsets to the control tablet and then copied to cloud storage for retrieval and analysis.

2.6 Fairness

We have scoped out our methodology to ensure fairness for all operators by adhering to the following procedures.

- Each network was tested concurrently to ensure that environmental conditions were the same for each operator.
- Identical handsets were used for each network: The Samsung Galaxy S6 Edge+ handsets.
- The handset position was rotated in the measurement equipment daily to evenly distribute the physical location of the handsets in the measurement backpacks.
- The SIM Cards were rotated between devices once a week to eliminate any differences that might occur from variations in individual handset performance throughout the testing.
- The measurement period was between 7am and 7pm each day, including weekends.

2.7 Test schedule

The measurement schedule was set up in the order shown in the table below. The test shown in bold fonts were used for internal testing purposes and were not included in our published metrics.

Data measurement schedule

HTTP Download (30s)	Max test duration 30s
	Max setup time 30s
Ping 32 bytes (5 pings)	**100ms Interval**

	Timeout 1s
	Max test duration 5s
Ping 800 bytes (10 pings)	**50ms Interval**
	Timeout 1s
	Max test duration 5s
Ping 32 bytes (25 pings)	10ms Interval
	Timeout 1s
	Max test duration 5s
PAUSE 15s	
HTTP Upload (15s)	Max test duration 15s
	Max setup time 30s
PAUSE 15s	
HTTP Browser mKepler	Max test duration 15s
HTTP Browser BBC	Max test duration 15s
HTTP Browser Amazon	Max test duration 15s
HTTP Browser YouTube	Max test duration 15s
PAUSE 15s	
HTTP Browser 'Big'	**Max test duration 15s**
PAUSE 15s	
YouTube	Display Duration 20s
	Max Duration 40s
	Connection timeout 25s
	Stream loss timeout 20s

Voice Measurement Schedule

Voice Quality and Statistics Test (POLQA)	Sample frequency 12.5s
	Call duration 90s
	Max call setup time 20s

PAUSE 15s

The 5 x 32 byte ping tests and the HTTP browser ‘Big’ tests were included for additional analysis but not published in our final report.

2.8 SIM cards

All of the operators we tested supplied SIM cards to Ofcom for use for the duration of our measurements. Due to the amount of data to be used during these tests (in the order of 6-7 GB per phone per day), the SIMs provided were standard consumer examples with the volumetric data caps removed.

The performance of the MNO-supplied SIM Cards were compared to the performance of consumer SIM Cards to ensure that that they were performing in the same way and had not been “optimised” for our testing. The consumer SIM cards were purchased on Ofcom’s behalf via a 3rd party and supplied on a twelve-month contract basis, to ensure that they were representative of available consumer tariffs.

2.9 Quality control: during measurement and post measurement

The following checks were conducted during testing to ensure the integrity of our results prior to processing:

- Prior to the start of testing each day, the arrangement of SIM Cards, Handsets and their physical location in the backpack were checked according to the predetermined schedule.

- The correct measurement schedule was loaded onto the devices.

- All testing was conducted by Ofcom engineers operating in pairs to ensure that the tests could be frequently observed via the control tablet regardless of whether the measurements were static, walking or driving.

- Markers were added to the generated data files to confirm what type of test was running and to denote whether the tests were being conducted indoors or outdoors.

- A manual log was maintained by the engineers in case clarification was required during our analysis.

- Replacement SIM Cards were held by the teams in case of problems with the removal of the volumetric cap on both Voice and Data usage and as a general backup.

Section 3

Data Processing

3.1 NQDI

Once the results were copied to Ofcom's data servers, it was then imported into our post-processing system, NQDI (Network Quality Data Investigator). NQDI is a Swissqual-supplied data analysis and report generating tool which enabled us to analyse and report on the data collected to compile the various metrics and statistics used in this report.

3.2 Data quality control

The measurement data from each city was retrieved from cloud storage to Ofcom's internal data server daily. This was then processed using NQDI and Excel to produce a report for each test conducted. The reports were then analysed by an external statistician and the final report produced.

After each day's measurement, the data were uploaded to local servers and then imported to the database using NQDI. A report was produced after the tests every day showing the number of samples collected for each test and for each technology.

Additional checks were made on the collected data prior to publication:

- The daily dashboard was used to check the number of samples collected for each test and for each operator was reasonable

- Checks were performed to see if enough samples were collected per city

- Any measurements that were aborted by the engineer or by the measurement equipment were identified and discarded

- Any data collected out of the predetermined core measurement hours (7am to 7pm) and outside the 10 km radius were filtered out

- Markers added to the data were double checked with paper logs for accurate position (indoor or outdoor) and state (either Walk/Drive/Static)

- The data was manually checked for missing data by looking for gaps in either date or time alongside the paper logs to ensure all the data is taken into account

- Each team was visited by a Quality Control assessor during their tests to ensure compliance with our predetermined processes

3.3 Weighting

In order to make comparisons between networks fair, it is desirable that the same number of readings for each network are taken within each city, indoor and outdoor, and by time of day – otherwise a network might benefit in the comparison by having more readings at quieter times of day, outdoor, or in easier locations. Obviously, the practicalities of fieldwork make this difficult to achieve, and so we weight the results to provide this equality. Put simply, if fewer readings are made in Cardiff outdoors during the weekend for O2 than for Vodafone, each Vodafone reading is weighted down by a factor which equalises this.

This is done after excluding unsuccessful and invalid tests, but not those that failed due to the network. It is done for all BB (Best Bearer) readings, regardless of the proportion that are carried on 4G as this is a measure of network capability rather than equalising the trial.

With the sample available for BB, this can be achieved with each city and network weighted to be 3.6% of the total (7 cities and 4 networks), and within this to have identical profiles by time of day and indoor/ outdoor. All significance tests take account of this weighting, which slightly reduces the accuracy compared to an unweighted sample, but is acceptable given the "fairness" it introduces into comparisons.

Analysis was also provided by distance (from centre), 0-4k or 4-10k, with comparisons of network, city and network within city required. For 0-4k the same approach could be used, but this was not possible for the 4-10k analysis as some cities did not have outdoor or weekend readings. Therefore, the 4-10k analysis was based on outdoor readings only, with time periods collapsed into three rather than four bands, by combining weekend and morning – based on the pattern of results overall. These are the most similar.

3.4 Averages

Ideally, we would use means to show an "average" score for each network, and conduct significance tests using the mean and its standard error. The problem is that such a test is only accurate if the underlying distribution is (close to) normal. In fact, if the sample is very skewed the mean can be considered a poor reflection of the "typical" customer experience, since we interpret it as the service the typical customer experiences. But, for example if 9 customers experienced a delay of 0.5 seconds and 1 a delay of 95.5 seconds, the mean score of 10 seconds would very poorly reflect "average" service.

With skewed distributions the median is considered a fairer comparison, as it is the value which 50% of customers do better than, and 50% do worse than. Using medians allows us to conduct significance tests without worrying about the underlying distribution, as follows:

- Take all the readings for the two subgroups, so for example all O2 and Vodafone readings in Cardiff
- Calculate the overall median
- Calculate the % for each subgroup (O2, Vodafone) that fall below the overall median – if the two networks performed the same, this would be 50% for each
- Test whether the observed %'s are consistent with this, using the standard test for comparing two percentages

To decide whether to use means or medians, we examined the distribution – if this were normal or close to normal, we would find one eleventh (9.1%) of the readings in each of the following ranges:

- More than 1.335 standard deviations less than the mean (R1)
- Between 1.335 and 0.908 standard deviations less than the mean (R2)
- Between 0.908 and 0.605 standard deviations less than the mean (R3)
- Between 0.605 and 0.349 standard deviations less than the mean (R4)
- Between 0.349 and 0.114 standard deviations less than the mean (R5)
- Between 0.114 less than and 0.114 standard deviations greater than the mean (R6)
- Between 0.114 and 0.349 standard deviations greater than the mean (R7)
- Between 0.349 and 0.605 standard deviations greater than the mean (R8)

- Between 0.605 and 0.908 standard deviations greater than the mean (R9)
- Between 0.908 and 1.335 standard deviations greater than the mean (R10)
- More than 1.335 standard deviations greater than the mean (R11)

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
HTTP Transfer Download	0%	14%	20%	14%	10%	8%	7%	6%	5%	6%	11%
HTTP Transfer Upload	5%	14%	13%	10%	10%	9%	10%	8%	4%	6%	12%
Call Statistics	3%	16%	9%	9%	10%	11%	14%	12%	6%	3%	7%
Voice Quality	10%	6%	5%	6%	8%	12%	16%	12%	10%	17%	0%

The following were not, and medians were used:

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
HTTP Browser mKepler	0%	0%	3%	26%	29%	22%	10%	3%	2%	1%	4%
HTTP Browser BBC	0%	6%	19%	17%	13%	11%	10%	8%	6%	4%	6%
HTTP Browser Amazon	0%	1%	13%	25%	22%	15%	9%	5%	3%	2%	5%
HTTP Browser Youtube	0%	0%	0%	25%	42%	15%	6%	3%	2%	2%	4%
StreamTimeToPic Data	0%	0%	41%	27%	4%	1%	1%	1%	0%	0%	25%
Ping	0%	0%	1%	20%	29%	32%	10%	3%	1%	1%	2%
Stream VMOS	14%	13%	1%	0%	2%	4%	1%	4%	61%	0%	0%
Stream Freeze	0%	65%	0%	0%	1%	5%	2%	0%	1%	12%	13%
Stream Jerkiness	0%	0%	63%	2%	1%	4%	0%	2%	1%	13%	12%

For the last three, the distribution is very skewed but the issue with using the median is that there is a peak value attracting over 50% of readings, so medians are often identical. The same test is used for these as for other metrics, but in this case it can mean that "differences" are marked as significant when the medians are actually identical. However, the test is still a valid one, as if the distributions were the same the difference would not be significant.

For video resolution, the test is of the percentage return of 720p, so neither mean nor median is necessary/ appropriate.

Compared to the previous phase, two metrics have changed: 'BBC browser' has moved from mean to median and 'voice' from median to mean. The latter certainly is a close call for both this phase and the previous phase of measurement, but for BBC Browser the shift in the distribution is noticeable