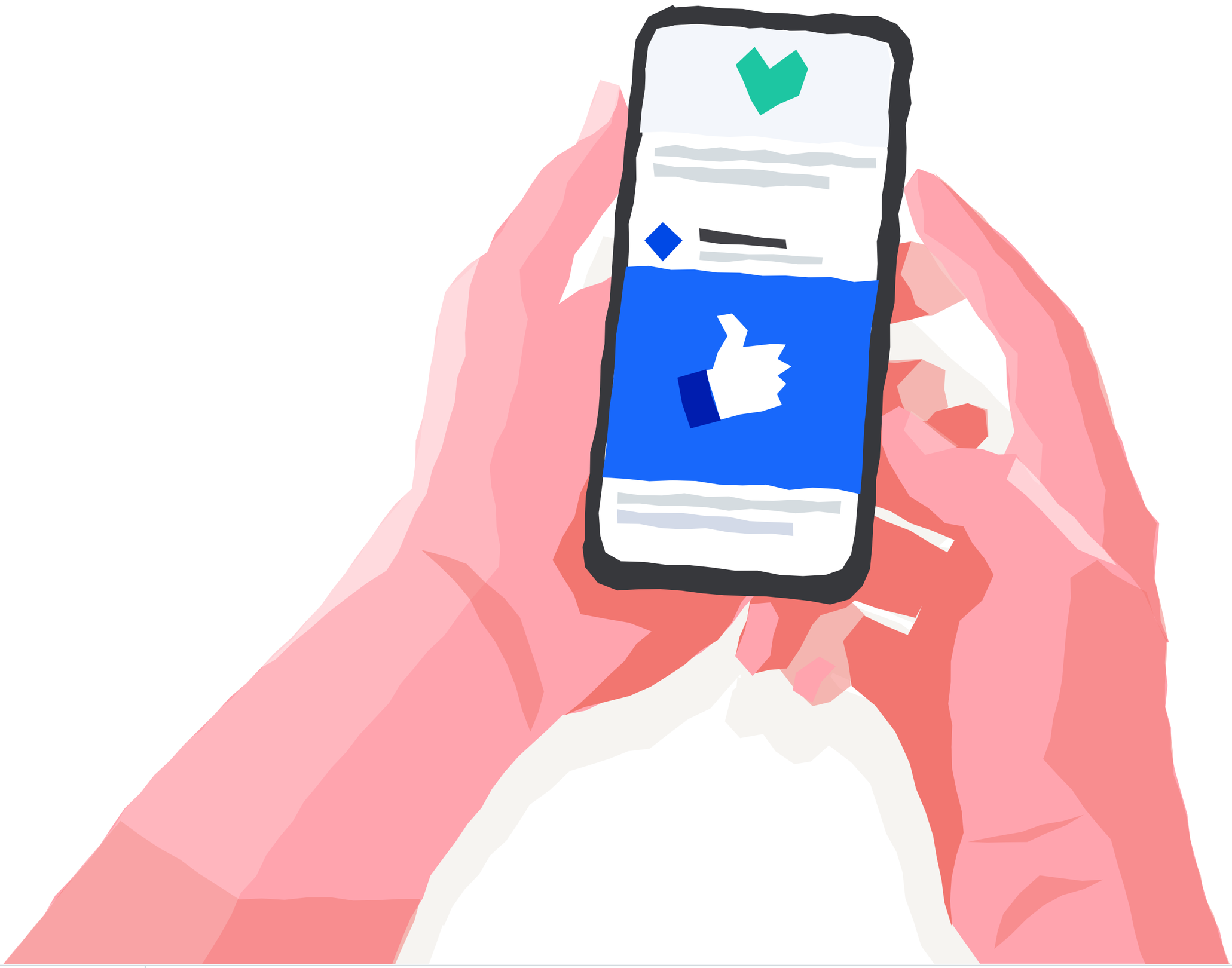
Measuring Broadband New Zealand

Data Dictionary, Summer Report, April 2021

In 2018, the Commerce Commission appointed SamKnows to measure New Zealand's internet performance. The programme, called Measuring Broadband New Zealand, gives internet users in New Zealand access to the SamKnows platform to measure the quality of their fixed-line internet. The aim of the programme is to increase transparency about actual in-home broadband performance and provide consumers with independent information about internet performance across different providers, plans, and technologies, to help them choose the best broadband for their homes. It will also encourage providers to improve and compete on their performance. This report provides an overview of the findings from data collected between 1st December and 31st December 2020.



Raw Data

Alongside the report ComCom also release the raw data and summary unit information used to produce the Summer 2021 report.

Two levels of data are included in this publication: **Raw Measurement Data** and **Per-Whitebox Summary Data**. More information on what is included in these files is outlined below.

Raw Measurement Data

This is the measurement data in its raw, unaggregated form. Only measurements that were used within the report have been included in this raw data package. Additionally, metadata fields which were not used in the report have been excluded (e.g. RSP name and product in specific instances).

The raw data is available in the ‘./raw_data’ directory, and a data dictionary describing the fields is included later in this document.

Per-Whitebox Summary Data

The measurements in the raw data are aggregated by Whitebox ID (also known as unit_id) as part of the data analysis process. The per-Whitebox data is far smaller, and therefore more accessible to third parties, than the raw data. It also includes additional derived fields which are used later in the analysis.

This summary data is calculated from the raw data using the statistical analysis tool R. The eventual aim is to release the R script used to create the charts along with the raw data so that interested parties can recreate the results. Due to the fact that certain metadata fields are excluded in the raw data, results in the report cannot be fully replicated with this data release.

A data dictionary describing the fields is included later in this document.

File listing

File	Description
./raw_data	
raw_download_tests.csv	Download speed test data
raw_upload_tests.csv	Upload speed test data
raw_latency_tests.csv	Latency and packet loss data
raw_netflix_tests.csv	Netflix data
raw_video_conferencing_tests.csv	Video conferencing data
./output	
report_charts.csv	Data behind the graphs which appear in the Summer report
unit_summary_statistics_download_upload_latency.csv	One line per Whitebox with download, upload and latency results
unit_summary_statistics_netflix.csv	One line per Whitebox with Netflix results
unit_summary_statistics_video_conferencing.csv	One line per Whitebox per video conferencing service

Raw Data

raw_download_tests.csv (Download speed)

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	Hostname of the test server.
download_mbps	decimal	Test speed in Mbps.
did_test_complete_successfully	boolean	Did the speed test complete successfully?
target_server_country	string	The country in which the test server is located.

raw_upload_tests.csv (Upload speed)

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	Hostname of the test server.
upload_mbps	decimal	Test speed in Mbps.
did_test_complete_successfully	boolean	Did the speed test complete successfully?
target_server_country	string	The country in which the test server is located.

raw_latency_tests.csv (Latency & Packet Loss)

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	Hostname of the test server.
latency_ms	decimal	The time for a round trip from Whitebox -> Server -> Whitebox.
num_successes	int	Number of packets which made a successful round trip.
num_failures	int	Number of packets which failed to make a round trip.
packet_loss_pct	decimal	Ratio of packets which did not make a successful round trip: failures divided by (successes + failures).
target_server_country	string	The country in which the test server is located.

Raw Data

raw_netflix_tests.csv (Netflix)

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
target	string	Hostname of the server assigned by Netflix to stream content.
bitrate_mbps	decimal	The bitrate that can be reliably streamed without stalls (in Mbps).
download_mbps	decimal	Download speed to the assigned Netflix server in Mbps.
did_test_encounter_stall	boolean	Did the test encounter a stall event?
did_test_complete_successfully	boolean	Did the test complete successfully?

raw_video_conferencing_tests.csv

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
dtime	datetime	The time of the test (UTC).
ddate	date	The date of the test.
is_during_peak_hour	boolean	Is the test in peak hour (7-11pm Mon - Fri)?
service	string	The video conferencing provider being tested.
region	string	The region of the server, as defined by the service provider.
latency_ms	decimal	The time taken to establish a connection with video conferencing services (in milliseconds).
did_test_complete_successfully	boolean	TRUE if any packets made a successful round trip, FALSE if not.
num_successes	int	Number of packets which made a successful round trip.
num_failures	int	Number of packets which failed to make a successful round trip.

Raw Data

unit_summary_statistics_download_upload_latency.csv

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
target_server_country	string	The country in which the test server is located.
trimmed_mean_download_mbps_24h	decimal	The 1% trimmed mean (average of the middle 98% of data) of download_mbps. Results where download_samples_24h is less than 5 are removed from the final dataset.
trimmed_mean_download_mbps_peak	decimal	The 1% trimmed mean (average of the middle 98% of data) of download_mbps - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE. Results where download_samples_peak is less than 5 are removed from the final dataset.
download_samples_24h	int	The number of download tests (count of rows in raw_download_tests.csv).
download_samples_peak	int	The number of download tests (count of rows in raw_download_tests.csv) - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE.
trimmed_mean_upload_mbps_24h	decimal	The 1% trimmed mean (average of the middle 98% of data) of upload_mbps. Results where upload_samples_24h is less than 5 are removed from the final dataset.
trimmed_mean_upload_mbps_peak	decimal	The 1% trimmed mean (average of the middle 98% of data) of upload_mbps - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE. Results where upload_samples_peak is less than 5 are removed from the final dataset.
upload_samples_24h	int	The number of upload tests (count of rows in raw_upload_tests.csv).
upload_samples_peak	int	The number of upload tests (count of rows in raw_upload_tests.csv) - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE.
trimmed_mean_latency_ms_24h	decimal	The 1% trimmed mean (average of the middle 98% of data) of latency_ms. Results where latency_samples_24h is less than 5 are removed from the final dataset.
trimmed_mean_latency_ms_peak	decimal	The 1% trimmed mean (average of the middle 98% of data) of latency_ms - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE. Results where latency_samples_peak is less than 5 are removed from the final dataset.
latency_samples_24h	int	The number of upload tests (count of rows in raw_latency_tests.csv).
latency_samples_peak	int	The number of upload tests (count of rows in raw_latency_tests.csv) - only considering tests during peak hours i.e. where is_during_peak_hour is TRUE.

Raw Data

unit_summary_statistics_netflix.csv

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
mean_netflix_download_mbps	decimal	Mean average of test results in Mbps.
netflix_samples	int	The number of times a Whitebox attempted to stream content from Netflix's CDN
max_concurrent_uhd_streams_old_classification	decimal	The number of simultaneous videos which could be streamed on average with traditional encoding. Assumes that an Ultra High Definition video will stream reliably with 15.6 Mbps of downstream bandwidth to Netflix's server.
max_concurrent_uhd_streams_new_classification	boolean	The number of simultaneous videos which could be streamed on average with Netflix's new variable bitrate shot-based encoding. Assumes that an Ultra High Definition video will stream reliably with 12 Mbps of downstream bandwidth to Netflix's server.

unit_summary_statistics_video_conferencing.csv

Field Name	Type	Description
unit_id	int	Unique identifier for an individual Whitebox.
service	string	The name of the video conferencing tested.
trimmed_mean_latency_ms	decimal	The 1% trimmed mean (average of the middle 98% of data) of hop_count. Results where video_conferencing_samples is less than 5 are removed from the final dataset.
video_conferencing_samples	int	The number of successful tests.