On the Validity of Internet Speed Test Tools and Broadband Measurement Programs

Jun Han
University of New South Wales
Sydney, Australia
jun.han3@student.unsw.edu.au

Minzhao Lyu
University of New South Wales
Sydney, Australia
minzhao.lyu@unsw.edu.au

Vijay Sivaraman
University of New South Wales
Sydney, Australia
vijay@unsw.edu.au

ABSTRACT

Speed is etched into our psyche – consumers like to buy broadband or mobile service from the "fastest" provider; there is a whole industry of speed testing services; governments invest millions of dollars in national speed monitoring programs; and Internet Service Providers (ISPs) use speed test rankings in their advertising. Given how strongly speed shapes consumer, industry, and government behaviour, one would expect speed testing and comparison methodologies to be thoroughly studied and documented; however, this is not so. In this paper, we first highlight how test conditions skew results. Using popular speed test tools, we highlight some of the factors – test duration, number of threads, congestion control algorithm, and server locations – that significantly impact outcomes. Speed results should therefore only be interpreted in the strict context of their test conditions, and any generalisations should be discouraged and discarded. Our second contribution shows that national monitoring programs that rank ISPs on speed are largely flawed. By analysing public data from Australia and the U.K., we show that sampling imbalances across ISPs in terms of the test conditions (*e.g.,* access technologies and locations) lead to biased results. Our study urges the community to restrain the emphasis placed on speed testing.

CCS CONCEPTS

• **Networks → Network measurement**;

KEYWORDS

Internet speed test, broadband measurement program

AINTEC'22, December 19–21, 2022, Hiroshima, Japan

ACM ISBN 978-1-4503-9981-4/22/12…$15.00
https://doi.org/10.1145/3570748.3570749

ACM Reference Format:
Jun Han, Minzhao Lyu, and Vijay Sivaraman. 2022. On the Validity of Internet Speed Test Tools and Broadband Measurement Programs. In *The 17th Asian Internet Engineering Conference (AINTEC'22), December 19–21, 2022, Hiroshima, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3570748.3570749

1 INTRODUCTION

Access network speeds have grown by several orders of magnitude – from Kilobits-per-second to Gigabits-per-second – over the past three decades. This has undoubtedly enabled users to enjoy streaming, gaming, conferencing, and other online activities at far superior levels of experience than ever before. However, beyond a certain limit, user perception of speed starts saturating [8], and higher speeds do not necessarily contribute to a better experience on most of the popular applications prevalent today. Nevertheless, the coupling between speed and experience is etched strongly into our psyche, and this blind faith leads consumers, governments, and the industry at-large to make decisions that are not always in society's best interests.

It is quite common for broadband and mobile users to do a speed test when they subscribe to a new service, are in a new connected environment, or encounter a performance issue. Popular speed test tools such as Ookla [21], M-Lab by Google [11], and Fast.com by Netflix [6], are freely available as browser utilities or smartphone applications – Ookla reports that over 43 billion speed tests have been conducted globally on its platform to-date [20]. Several governments run nationally funded broadband speed monitoring programs, with the intent of informing citizens how Internet Service providers (ISPs) compare on speed delivered versus promised in the plan. These include UK's Broadband Speed Research program by the Office of Communications (Ofcom) [18], USA's Measuring Broadband America program by the Federal Communications Commission (FCC) [32], and the Australian Competition & Consumer Commission (ACCC)'s Measuring Broadband Australia program [28]. These programs deploy measurement robots into volunteers' homes that perform periodic speed tests, and the averaged results published monthly or quarterly are used as a basis for ranking ISP performance. As a visual example shown in Fig. 1,

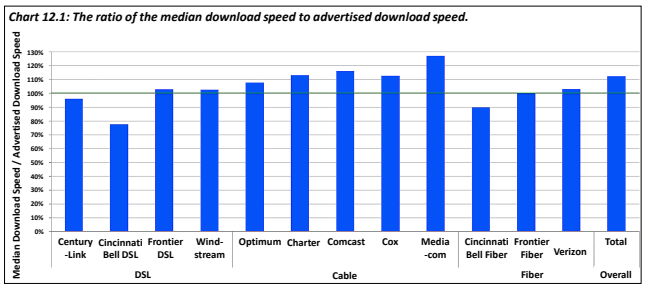


Figure 1: The ratio of measured median download speed to advertised download speed per ISP, served as a basis of ISP ranking, taken from the FCC MBA Eleventh Report (U.S.), Dec 2021 [29].

FCC US published its yearly report to rank the broadband performance of different ISPs and access technologies by the ratio of the user received (median) speed over the advertised plan speed.

Given the significant role of speed test outcomes on the individual's purchasing decisions, ISP's advertising campaigns, and nation's monitoring programs, one would expect their reliability and robustness to be firmly established. Unfortunately this is not so. Prior works [4, 7, 13, 14, 23, 31] have shown that aspects of the testing environment (*e.g.,* user device type, router vs device-based measurement, ISP link quality, and server capacity) can impact the accuracy of speed test results. Further, many speed test tools can be configured to adjust duration, number of parallel connections, and server locations, with no guidance provided to users on how to select them appropriately [2]. Lastly, it can be very difficult to detect and correct for differences in speed test conditions, which can lead to incorrect inferences [5].

In this paper we make two contributions. For our first contribution (§3) we systematically evaluate the reliability of speed measurements as various aspects of the testing conditions change. We consider some popular speed test tools in the market, and conduct experiments that vary test parameters such as number of threads, test duration, congestion control algorithm, and server location. We demonstrate that test results can vary significantly based on parameter choices. Two test results are therefore comparable only if the testing conditions are identical, and even then their comparison rank can change under different testing conditions. Speed test results should therefore be interpreted with a high degree of caution.

For our second contribution (§4), we analyse raw datasets made publicly available from national broadband measurement programs of the U.K. and Australia, and show that sampling is imbalanced across ISPs in aspects such as access technologies and locations. These lead to inaccurate inferences as ISP rankings become artefacts of test conditions rather than network superiority. This raises questions on whether citizens are being correctly informed on broadband competition, and that too using taxpayer funds.

The rest of this paper is organized as follows. Background on Internet speed tests and related works evaluating their performance are summarized in §2; our first contribution that evaluates speed measurement tools of various configurable test conditions is presented in §3; our second contribution that highlights the problem of imbalanced sampling in current broadband measurement programs is discussed in §4; and this paper is concluded in §5.

2 BACKGROUND AND RELATED WORK

While network performance could be interpreted by various metrics such as round-trip time, jitter, or packet loss, network speed (also known as throughput), is among the most important factors that decide a user's experience in broadband networks [2]. Individual customers, content providers (*e.g.,* Netflix [15]), and governmental organizations use this metric to measure the service quality provided by various ISPs, which could serve as a reference for their decisions on subscription, investment, and policy-making [7].

A speed test tool, in general, measures the maximum throughput between a client and the test server that is carried by one or multiple concurrent TCP flows (*i.e.,* threads) that last for a certain duration. The test server, number of threads, and duration are either fixed or configurable to users, depending on the certain tools used. For example, among the four tools discussed in this paper, Ookla, Fast, and iPerf3 allow clients to configure at least one parameter, whereas SamKnows, a measurement provider mostly focusing on business users, has all test parameters fixed in its tool.

At the national level, broadband measurement programs have been launched in many countries to monitor the quality of network services provided by various ISPs, under different access technologies, across different regions, etc. Five national broadband speed measurement programs that will be discussed later in §4 [16, 18, 25, 28, 32] all use SamKnows as their test provider that conducts broadband measurement from distributed vantage points. Most of the programs publish their measurement datasets periodically (*e.g.,* per annual) along with an ISP ranking report that is predominantly on the median or the mean values of the measured speeds.

There are several research works that focus on evaluating or increasing performance of speed test tools [2, 4, 7, 13, 14, 23, 31] that inspire our study. S. Bauer *et at.* [2] discussed how various speed test tools give different measurement results; N. Feamster *et al.* [7] pointed out that the accuracy of speed test may not actually reflect the quality of an ISP network, as it highly depends on the environment setup such as user device hardware, connection type (*i.e.,* wireless or wired), operating

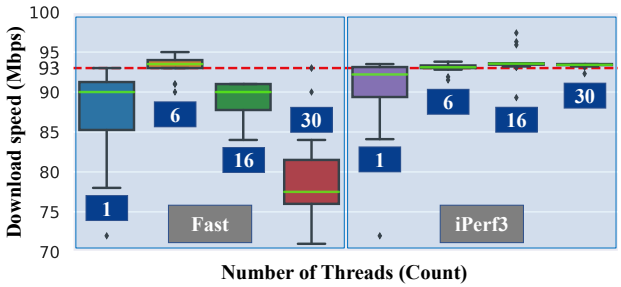


Figure 2: Speed test results of Fast and iPerf3 with different numbers of threads ranging from 1 to 30.

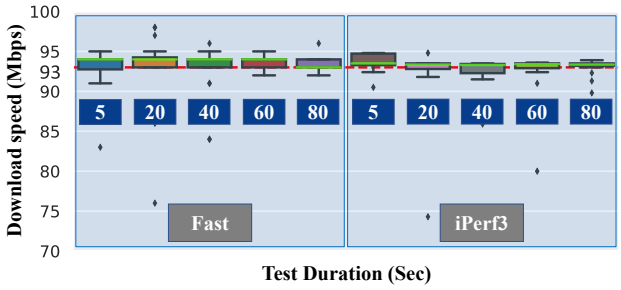


Figure 3: Speed test results of Fast and iPerf3 with different test durations ranging from 5 to 80 seconds.

systems, and browser version; and K. Macmillan *et al.* [13] studied how client-server latency, client device, and access medium affect speed test results. The works in [4], [31], and [14] proposed specialised speed test methods/tools to achieve better performance for measuring broadband, TOR, and cloud platform, respectively.

3 EVALUATING INTERNET SPEED TESTING TOOLS

To understand how individual speed test results could be biased by measurement configurations, in this section, we empirically explore the impact of different configurable parameters on three popular speed testing tools under a well-controlled lab environment (§3.1). Their result variations are reported as dependencies on the number of threads (§3.2), testing duration (§3.3), server locations (§3.4), and TCP congestion control algorithms (TCP CCAs) (§3.5).

3.1 Experiment Setup

We select three popular speed testing tools (namely Ookla [21], Fast [6], and iPerf3 [9]) on the market that are freely accessible by users via either browser or mobile applications. Fast and iPerf3 enable their users to configure the number of concurrent measurement threads and test duration, Ookla allows users to select their test destination servers at different geolocations, and iPerf3 allows users to configure their TCP CCAs. Thus, in what follows, performance impacts caused by the four configurable measurement parameters, *i.e.,* the number of threads, test duration, server location, and TCP CCA on their respective tools are thoroughly evaluated.

To avoid the possible impacts of other factors such as device specification, network capability, and software version, which have already been studied in prior works, in our experiment, active measurements by the three speed testing tools were conducted by a Chrome browser on the same PC running Window10 OS (*i.e.,* user). The test laptop is connected to our controlled lab network with its advertised downstream speed of 100 Mbps. Our measurement PC is the only active device under the lab network. Therefore, the performance variations in test results are solely introduced by the changes in our configurable parameters.

3.2 Number of Concurrent Threads

Two of the studied speed test tools, namely Fast and iPerf3, allow users to configure the number of active threads connected to the test servers. To understand the result variations caused by this parameter, we perform measurements using the two test tools with a different number of threads, while other parameters remain identical. Thirty speed test results are collected for thread counts as 1, 6, 16, and 30 using the respective tools. The distribution of results is visually shown as a box plot in Fig. 2. We note that 30 is the maximum thread count of Fast, while 6 and 16 are practical choices suggested by the speed test community [2, 22].

Intuitively, we expect to have a more accurate result (*i.e.,* closer to the advertised speed) with a larger number of measurement threads, as it could effectively smooth the random variation of each single thread. This hypothesis is aligned with our empirical results for iPerf3 (the right half side of Fig. 2), where the variation of test results under a higher number of threads (*i.e.,* 6, 16, and 30), as indicated by the span of each box, are significantly reduced. Therefore, a test with higher thread counts is more likely to reach its actual received speed of around 93 Mbps. However, such conclusion cannot be drawn universally. The performance of some speed testing tools (*e.g.,* Fast) reaches their best cases at a certain number of threads, and starts falling (dramatically) after that crest point. According to the left half side of Fig. 2, Fast has its minimum empirical result variations with 6 measurement threads, where the median value is also very close to the ground-truth speed (*i.e.,* 93 Mbps). Whereas the testing results become worse (*i.e.,* larger variations and smaller median values) with 16 and 30 measurement threads.

Key Takeaway: Based on the above observations, we note that, while one concurrent measurement thread can

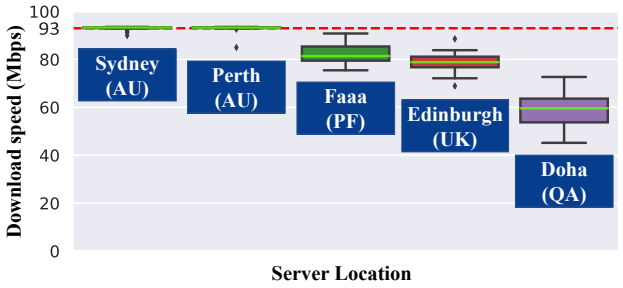


Figure 4: Speed test results of Ookla with different server locations.

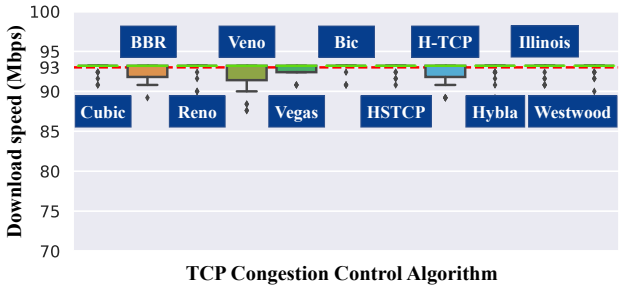


Figure 5: Speed test results of iPerf3 with different TCP congestion control algorithms.

hardly give an accurate result, increasing the thread count to a high value does not always give better performance, such as our measurement results of Fast in Fig. 3. A user would have to select the optimal number of concurrent measurement threads (*i.e.,* the crest point) to achieve an accurate estimation of her/his received network speed.

3.3 Test Duration

Now we study how the measurement performance changes with respect to different configurations on test duration, indicating how long a speed test lasts. This parameter is configurable on both Fast and iPerf3 tools. The server locations are fixed and not configurable by both tools. We also fix the number of threads to 6, the near-optimal value for both Fast and iPerf3 as identified above, so that the test duration is the only variable, ranging from 5 to 80 seconds.

As our initial hypothesis, the increase in test duration would inevitably lead to a more reliable measurement result, as the measured speed is likely to be converged to its stable state over a considerable duration. This assumption is confirmed by our empirical results shown in Fig. 3. The left half figure shows the distribution of measured speed with different test duration via Fast, and the right half figure depicts the results using iPerf3. It is clear that, in general, both tools exhibit better performance (*i.e.,* minimized variations as indicated by the box length) with a longer test duration. However, the decrease of variation becomes less significant when the test duration is longer than 20 seconds, as revealed by the green, red, and purple boxes in Fig. 3, whereas iPerf3 keeps getting better performance until reaching the maximum duration (*i.e.,* 80 seconds).

Key Takeaway: For a user measuring its received network speed, increasing test duration would likely reach a more accurate result. Depending on the measurement setup, there is an optimal test duration with which a user could have her/his best-expected measurement result without spending unnecessary time, which could be longer than one minute. However, to the best of our knowledge, such optimal duration is not documented by the existing speed test tools on the market, which makes it non-trivial for users to figure the value out.

3.4 Server Locations

Now we study the performance impact caused by different destination server locations. Of the three speed testing tools, Fast automatically selects the closest destination server to the user, while iPerf3 and Ookla allow their user to choose her/his test servers by location. iPerf3 has a limited number of public servers around the globe, while Ookla has a comparatively larger bank of servers to choose from. We now conduct our measurements using the Ookla tool by select servers operated by Vodafone at five different geolocations from three countries (*i.e.,* Sydney (AU), Perth (AU), Faaa (PF), Edinburgh (UK) and Doha (QA)). During measurement, Ookla uses a fixed 16 threads and test duration that could not be configured by a user. Hence, server location is the only variable in this set of measurements using Ookla.

Intuitively speaking, the measurement results would be quite stable (*i.e.,* close to the actual value with a very low chance of having a large deviation) when the destination server is close to the client, as there are likely to be fewer hops on the routing path, each could introduce measurement variation to some extend. The distribution of measurement results toward each of the five server locations is presented as box plots in Fig. 4.

From the box plot, it is quite clear that measurement results toward servers at close locations (*i.e.,* Sydney (AU) and Perth (AU)) are with quite small variations (indicated by the box length) and median values approaching the ground-truth speed. Whereas the results of further server locations not only have larger variations, but also median values that are far away from the actual received speed (*i.e.,* 93 Mbps).

Key Takeaway: As expected, long-distance (*i.e.,* many hops on the routing path) between a user and the test server often introduces additional biases that could significantly

Table 1: Availability of test condition labels of the publicly available datasets from four national broadband speed test programs in the year of 2021.

Datasets	ISP/RSP	Geolocation	Speed Tier/Plan/Package	Access Technology
ACCC AU [24]	Provided	State/Territory Level	Provided	Provided
Ofcom UK [19]	Provided	Urban/Rural & Kingdom Level	Provided for partial ISPs	Provided for partial ISPs
FCC US [33]	Not Provided	Not Directly Provided	Not Directly Provided	Not Provided
CC NZ [17]	Not Provided	Not Directly Provided	Not Directly Provided	Not Provided

deviate the measured results from their ground-truth speed. In practice, speed testing tools often have limited choices of destination servers (*i.e.,* five servers in our empirical study), which may not contain the optimal location for a user. Therefore, without a properly chosen destination server that could be impractical under certain circumstances, a measured result may not necessarily represent the actual network speed a user receives.

3.5 TCP Congestion Control Algorithms

We now demonstrate how different TCP congestion control algorithms (CCAs) could skew test results. For the three test tools considered, Ookla and Fast have fixed CCA, while iPerf3 leave this choice to users. Thus, we now study the performance biases by iPerf3 with different CCAs including Cubic, BBR, Reno, Veno, Vegas, Bic, HSTCP, H-TCP, Hybla, Illinois, and Westwood. Other parameters are fixed so that we have a controlled environment for our tests.

Different CCAs have their own performance metric(s) to maximise during operation. For example, BBR [3] aims to achieve optimal bandwidth and round trip time and Vegas [12] focuses on packet delays. Thus, not surprisingly, the TCP-based test results with varied CCAs are inevitably biased due to their different underlying mechanisms. As shown in Fig. 5, among the 11 CCAs, seven of them have relatively stable results around 93 Mbps with small variations (*i.e.,* small box length), whereas the rest four CCAs (*i.e.,* BBR, Veno, Vegas, H-TCP) exhibit observable variations (*i.e.,* with box lengths as around 1, 2, 0.5, and 1 Mbps, respectively) in their test results.

Key Takeaway: Different TCP CCAs could introduce non-negligible biases into the speed test results, though it is not configurable nor acknowledged to users in most of the speed testing tools. Optimally selection of CCA on the client side seems to be a necessary but not trivial step for users to achieve a reliable measurement of their actual received speed.

4 EVALUATING NATIONAL BROADBAND MEASUREMENT PROGRAMS

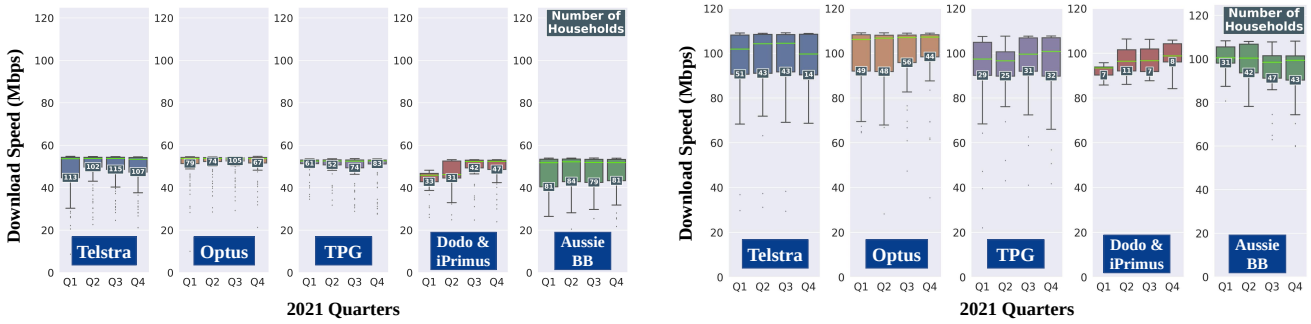
Having seen that several factors can bias speed test results, we now analyse data from multiple national broadband measurement programs to estimate the prevalence of biases and impact on inferences. The datasets are described (§4.1), inconsistencies across ISPs explored (§4.2), and biases in access technologies (§4.3) and geographies (§4.4) revealed.

4.1 Comparing the Published Datasets of Four National Programs

We now look at the national speed test programs for five countries including the U.S., U.K., Canada, Australia, and New Zealand, as discussed in §2. First, four countries (*i.e.,* the U.S., U.K., Australia, and New Zealand) have their measurement results periodically published in comma-separated format (.csv) files with analytical reports, whereas the Canada program only makes its ranking reports publicly available without a measurement dataset. Therefore, our conclusions drawn in this section do not cover Canada's CRTC program.

For the four programs, we analysed their published datasets for the entire year of 2021 to understand whether their measured speeds from distributed vantage points are sufficient to represent the broadband performance of the respective nation or not. It is worth noting that all the four programs employ the same speed test utility, SamKnows, a commercial speed test tool developed for both consumers and business users [26]. The choices of their configurable parameters during each test are not documented, therefore, as already discussed in §3, the accuracy of each test run may not be optimally tuned and consistent.

Notably, test conditions (*e.g.,* ISP, client geolocation, speed tier, or access technology) of each measurement results are not always sufficiently labelled in the four datasets, which are extremely important for post-hoc analysis to give fine-grained and actionable suggestions, *e.g.,* clients under which access technology are likely to receive worse speeds from their subscribed plans. A detailed category of test conditions and their availability in each of the four datasets are provided in Table 1. We could see from the table that, the ACCC AU dataset has quite comprehensive coverage of all test conditions. The Ofcom UK datasets have good labels of ISP and geolocations, however, it does not include all speed tiers and access technologies for all ISPs, *e.g.,* 1000 Mbps plan provided by Plusnet, one of the U.K.'s most popular ISP, has not been covered by the lastest UK dataset. As for the FCC US and the CC NZ datasets, they do not have clear labels on the most of test conditions listed in Table 1.



(a) 50/20 Mbps tier.

(b) 100/40 Mbps tier.

Figure 6: Measured download speed per household across different ISPs in ACCC AU 2021 dataset.

In what follows, to convey our key insights concisely, we mainly focus on our analysis of the two datasets (*i.e.,* ACCC AU and Ofcom UK) that have good coverage of test conditions, which are still identified to have inconsistency due to unbalanced measurement samples.

4.2 Measurement Inconsistency across ISPs

We first look at the performance inconsistency of measurement results with different ISP labels. The ACCC AU dataset provides the measured download speed from each client household (*i.e.,* measurement vantage point) under different ISPs, the five major ones among which are Telstra, Optus, TPG, Dodo & iPrimus, and Aussie Broadband. The median (or mean in the AU dataset) value of each ISP is often used as an indicator of the quality of services provided to their customers [27]. To understand whether such median values are sufficient to rank ISPs, in Fig. 6, we visualize the distribution of measured speed for each ISP during the four quarters ending in Feb, May, Sep, and Dec. Two speed plans (*i.e.,* 50/20 Mbps and 100/40 Mbps tiers) are considered separately in Fig. 6(a), and Fig. 6(b), respectively.

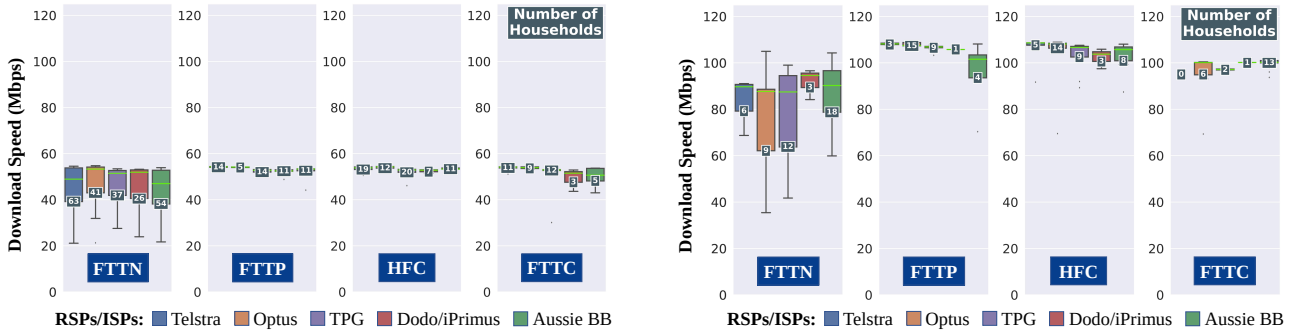
First, although the median values (shown as the green lines in each box) seem consistent for an ISP in the majority of the boxes over time, the large and inconsistent variations (depicted by the length of each box) within and across each ISP weaken the indicative power solely by the median. To be specific, in both of the box plots, Fig 6(a) and Fig. 6(b), the median values of boxes belonging to the same ISP have quite similar values in most of the cases. For example, for the 50/20 Mbps speed tier, all four boxes for Telstra have their median values around 53 Mbps, while the value for Optus, TPG, and Aussie Broadband is quite stable around 53 Mbps, 52 Mbps, and 51 Mbps, respectively. There is also an exception for Dodo & iPrimus with the 50/20 Mbps plan, for which the median values stay around 46 Mbps in the first two quarters and increase to 52 Mbps in Q3 and Q4. In the meantime, we could also observe that the variation of results (*i.e.,* size of each box) is quite diversified not only across various ISPs, but also within each ISP over time. A very obvious example is for Aussie Broadband in Fig. 6(b), where the median values of the four boxes are all close to 100 Mbps but the upper/low bounds and lengths of the boxes are clearly different. Similar observations are also obtained from the other three measurement programs, which are not explicitly presented here.

Key takeaway: Regardless of the seemingly consistent median values of each ISP that are predominately used as an indicator of broadband performance, the variation of measurement results also has quite diversified values but is not fully considered by the current ranking reports. Such differences in measurement variations might reflect the uneven service quality provided to users by an ISP, but also could be due to unbalanced selections of sample households that participated in the measurement program, which will be discussed next.

4.3 Measurement Inconsistency across Access Technologies

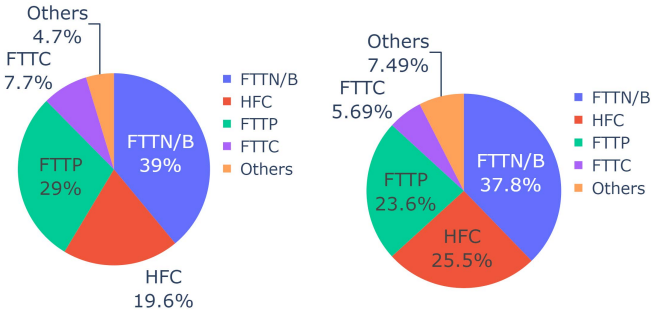
We now analyse the consistency of measurement results for four major broadband access technologies, including Fibre to the Node (FTTN), Fibre to the Premises (FTTP), Hybrid Fibre Coaxial (HFC) and Fibre to the Curb (FTTC). Ranking reports on those access technologies could serve as an important reference for infrastructure investments by ISPs or government departments. The distribution of measured download speed per household are grouped by their access technologies in Fig. 7, under 50/20 Mbps plan (Fig. 7(a)) and 100/40 Mbps plan (Fig. 7(b)), respectively. Each box contains measurement results of all households under the same ISP.

Comparing the medians and variations of each box in Fig. 7, it is not surprising to reach a conclusion that FTTP and HFC have the best performance under both 50/20 Mbps



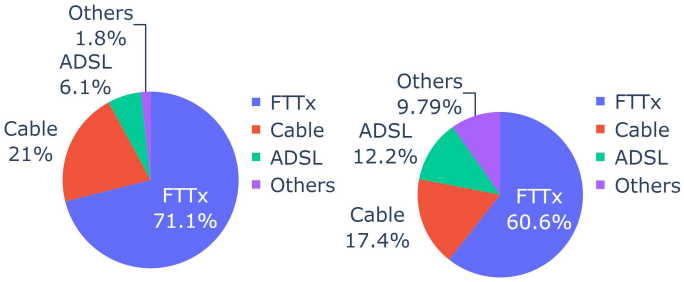
(a) 50/20 Mbps tier. (b) 100/40 Mbps tier.

Figure 7: Measured download speed per household with different access technologies during the fourth quarter of 2021 in the ACCC AU dataset.



(a) Measurement percentage. (b) Population coverage.

Figure 8: Composition of (a) measurement sample sizes and (b) the number of actual households on each type of access technologies during the fourth quarter of 2021 in ACCC AU dataset [1].



(a) Measurement percentage. (b) Population coverage.

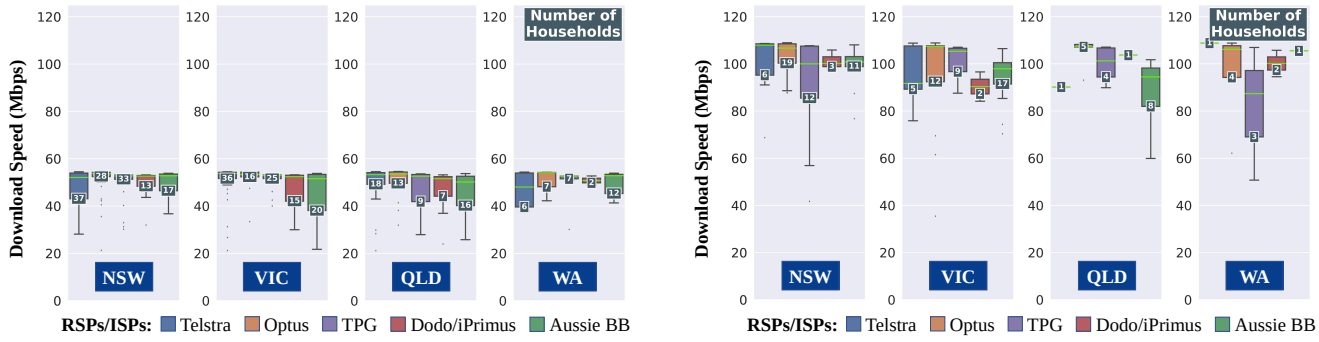
Figure 9: Composition of (a) measurement sample sizes and (b) the number of actual households on each type of access technologies during the fourth quarter of 2021 in Ofcom UK dataset [10, 30].

and 100/40 Mbps plans, not only for their higher median speed value than others but also for their more consistent performance (*i.e.,* small variations of the respective box plots). More importantly, FTTN has the worst performance in both of the metrics. Therefore, it seems quite necessary for all ISPs to improve their service quality for clients on FTTN.

However, such conclusions drawn from both median values and deviations are likely to be biased by the selected vantage points (measured household), which may not fully represent the true nature of users that could have diverse network profiles. In Fig. 7, the sample sizes (*i.e.,* household) included in each box are displayed on their bottom edges. A straightforward observation could be made that FTTN has the largest sample sizes compared with other technologies (*e.g.,* more than 20 households for each ISP under 50/20 Mbps plan in Fig. 7(a)), whereas other technology types often have less than 10 households in each box. Moreover, the sample size for FTTC provided by Telstra under the 100/40 Mbps plan is absolute zero, resulting in poor coverage of customers under this category.

To further understand the unbalanced sample sizes for each access technology type, we show the compositions of sample sizes and the actual number of Australian households on each type of access technology as pie charts in Fig. 8. Ideally, to reach a fair conclusion, the selected households in the measurement program should be evenly distributed across each technology type – a nearly perfect match of Fig. 8(a) and Fig. 8(b). However, due to small sample sizes, the current sampling methods do not achieve this expectation, and thus, could hardly give reliable references. For example, AussieBB (Aussie Broadband) is heavily oversampled on FTTN compared to Optus, especially for the 100/40 Mbps tier shown in Fig. 7(b), which could possibly explain why it has a lower overall rank.

Similar observations are also obtained from other measurement programs considered in this paper. For example, Fig. 9 shows the unbalanced sample sizes of three major access technologies (*i.e.,* FTTx, Cable, and ADSL) in the Ofcom UK dataset (Fig. 9(a)), which is clearly not matched with the actual population of the respective technologies shown in Fig. 9(b).



(a) 50/20 Mbps tier.

(b) 100/40 Mbps tier.

Figure 10: Measured download speed per household with different geolocations during the fourth quarter of 2021 in the ACCC AU dataset.

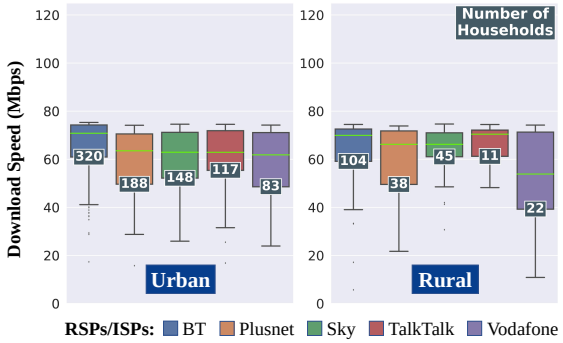


Figure 11: Measured download speed per household in urban or rural areas during Mar 2021 under 76/20 Mbps speed tier in Ofcom UK dataset.

Key Takeaway: Different access technology by nature could have significant impacts on the actual speed received by a user, which is partially reflected by the current measurement results. However, given the relatively small and unbalanced sample sizes with respect to their represented populations under each category, conclusions drawn from the results will be inevitably biased and thus not highly trustable when used as a reference in making critical decisions like infrastructure investment.

4.4 Measurement Inconsistency across Geolocations

Apart from the access technologies that are just discussed, possible result biases due to inconsistent sample sizes are also identified when considering geolocations. Now we discuss our findings by taking the ACCC AU dataset as a representative example, while similar observations are found in other datasets as well. First, of the eight states or territories in Australia, four of them (*i.e.,* NT, ACT, SA and TAS) have a negligible amount of samples to be properly analysed. The measurement results of the other four regions (*i.e.,* NSW, VIC, QLD, and WA) are also severely impacted by their unbalanced sample sizes, which will be explained as follows.

Considering the box plots of measured download speed per household in different regions, as shown in Fig. 10, we could see that many of the ISPs do not have a sufficient amount of sample size compared to others. For example, Telstra, a large ISP in Australia, only have 6 measured household under 100/40 Mbps speed tier in NSW, a state with the most population in Australia. Although it has a better median value (*i.e.,* 109 Mbps) from this very limited sample size when compared with another large provider TPG, which has 12 sample households, it is very difficult to reach a defensible conclusion that Telstra outperforms TPG at 100/40 Mbps plan in NSW due to their unbalanced sample sizes.

In addition, when comparing the broadband network performance across different regions in the country, it would be unfair to reach any conclusion from the very limited number of sample sizes, especially in WA and QLD, where the selected household under the high-speed plan (*i.e.,* 100/40 Mbps) are mostly less than 5 for a given ISP, as shown in Fig. 10(b).

To further strengthen our findings, we now discuss an interesting observation from the Ofcom UK dataset that is quite counter-intuitive. The UK dataset provides a label for each line of measured results to indicate whether the participating household is in an urban or rural area. We show the per household values in Fig. 11. From the median speeds and variations of each box, households in rural areas seem to have better network quality than those in urban regions like London and Manchester. This observation is particularly clear for three ISPs (*i.e.,* Plusnet, Sky, and TalkTalk), whose median speeds are about 10 Mbps higher in rural than urban regions. While this conclusion looks very promising as residents in the countryside could enjoy equivalent or even better broadband networks than people in capital cities, it is also quite dubious as the sample sizes under the urban

categories are more than five times larger than those of the rural ones. Therefore, rural households with poor broadband speeds might not be fully represented by the current measurement samples.

5 CONCLUSION

This paper has evaluated the validity of speed test tools and national-level speed measurement programs. We first designed experiments with three popular speed test tools and showed how different configurable speed test parameters or conditions such as test duration, number of threads, congestion control algorithm, and server locations can have a notable impact on speed test results. Then by studying and analysing national broadband measurement programs, mainly from Australia and the U.K., we demonstrated how their measurement datasets are biased by unbalanced sample sizes of different ISPs, access technologies, and geolocations. Our study revealed the reliability issues of current speed test methods and suggests both users and governmental agencies only interpret a test result with acknowledgment of its conditions and employ a fair and robust sampling strategy in large measurement programs.

REFERENCES

[1] Sam Baran. 2020. FTTC: What is NBN Fibre to the Curb? https://www.finder.com.au/fttc-nbn. (Jul 2020). Accessed: 2022-08-01.

[2] Steven Bauer, David Clark, and William Lehr. 2010. *Understanding Broadband Speed Measurements.* Technical Report.

[3] Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. 2016. BBR: Congestion-Based Congestion Control. *ACM Queue* 14, September-October (2016), 20 – 53.

[4] Marshini Chetty, David Haslem, Andrew Baird, Ugochi Ofoha, Bethany Sumner, and Rebecca Grinter. 2011. Why is My Internet Slow? Making Network Speeds Visible. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11).* Association for Computing Machinery, New York, NY, USA, 1889–1898.

[5] Xiaohong Deng, Yun Feng, Thanchanok Sutjarittham, Hassan Habibi Gharakheili, Blanca Gallego, and Vijay Sivaraman. 2021. Comparing Broadband ISP Performance using Big Data from M-Lab. *CoRR* abs/2101.09795 (2021).

[6] Fast. 2022. The Fast Speed Test (Powered by Netflix). https://fast.com/. (2022). Accessed: 2022-04-22.

[7] Nick Feamster and Jason Livingood. 2020. Measuring Internet Speed: Current Challenges and Future Recommendations. *Commun. ACM* (Nov 2020).

[8] Broadband Forum. 2021. An economic argument for moving away from Mbps. https://bit.ly/3E4ydsv. (Jul 2021). Accessed: 2022-08-01.

[9] IPerf. 2022. The iPerf3 Speed Test. https://iperf.fr/. (2022). Accessed: 2022-01-28.

[10] Ispreview. 2022. Summary - UK Fixed Line Broadband Statistics. https://www.ispreview.co.uk/broadband.shtml. (2022). Accessed: 2022-08-01.

[11] Measurement Lab. 2022. Measurement Lab Speed Test. https://bit.ly/3P4IIhB/. (2022). Accessed: 2022-04-22.

[12] Steven Low, Larry Peterson, and Limin Wang. 2000. Understanding TCP Vegas: Theory and Practice. (Mar 2000).

[13] Kyle MacMillan, Tarun Mangla, Marc Richardson, and Nick Feamster. 2022. Best Practices for Collecting Speed Test Data. (Aug 2022).

[14] Ricky K. P. Mok, Hongyu Zou, Rui Yang, Tom Koch, Ethan Katz-Bassett, and K C Claffy. 2021. Measuring the Network Performance of Google Cloud Platform. In *Proceedings of the 21st ACM Internet Measurement Conference.* Virtual Event.

[15] Netflix. 2022. Netflix. https://www.netflix.com. (2022). Accessed: 2022-08-01.

[16] The Commerce Commission (NZ). 2022. The Commerce Commission (NZ): Measuring Broadband New Zealand Program. https://bit.ly/2OHQLlX. (Apr 2022). Accessed: 2022-06-01.

[17] The Commerce Commission (NZ). 2022. Reports from Measuring Broadband New Zealand. https://bit.ly/3U9VgYy. (Aug 2022). Accessed: 2022-04-22.

[18] The Office of Communications (U.K.). 2021. The Office of Communications (U.K.): Broadband Speeds Research Program. https://bit.ly/3UmdWUQ. (Sep 2021). Accessed: 2022-04-22.

[19] The Office of Communications (U.K.). 2021. UK home broadband performance, measurement period March 2021. https://bit.ly/3U7A0mn. (Sep 2021). Accessed: 2022-04-22.

[20] Ookla. 2022. Ookla Speedtest: About. https://www.speedtest.net/about. (2022). Accessed: 2022-04-22.

[21] Ookla. 2022. The Ookla (speedtest.net) Speed Test. https://www.speedtest.net/. (2022). Accessed: 2022-04-22.

[22] Carin Overturf. 2019. Ookla: How does Speedtest measure my network speeds? https://bit.ly/3QadwyQ. (2019). Accessed: 2022-08-01.

[23] Udit Paul, Jiamo Liu, Vivek Adarsh, Mengyang Gu, Arpit Gupta, and Elizabeth M. Belding. 2021. Characterizing Performance Inequity Across U.S. Ookla Speedtest Users. *CoRR* abs/2110.12038 (2021).

[24] The Australian National Open Data Portal. 2022. The Australian National Open Data Portal: The ACCC Datasets. https://data.gov.au/data/organization/accc. (Aug 2022). Accessed: 2022-04-22.

[25] The Canadian Radio-television and Telecommunications Commission (CA). 2020. The Canadian Radio-television and Telecommunications Commission (CA): Measuring Broadband Canada Project. https://crtc.gc.ca/eng/internet/proj.htm. (Sep 2020). Accessed: 2022-08-01.

[26] SamKnows. 2022. SamKnows Offical Website. https://samknows.com/. (2022). Accessed: 2022-04-22.

[27] The Australian Competition & Consumer Commission (AU). 2022. *Measuring Broadband Australia Report 18, August 2022.* Technical Report. https://bit.ly/3E4GJrv

[28] The Australian Competition & Consumer Commission (AU). 2022. The Australian Competition & Consumer Commission (AU): Measuring Broadband Australia Program. https://bit.ly/3DZhXYL. (Aug 2022). Accessed: 2022-04-22.

[29] The Federal Communications Commission (U.S.). 2021. *Measuring Fixed Broadband - Eleventh Report.* Technical Report. https://bit.ly/3FNNkrp

[30] The Office of Communications (U.K.). 2021. *Connected Nations 2021.* Technical Report. https://bit.ly/3NBWjxU

[31] Matthew Traudt, Rob Jansen, and Aaron Johnson. 2021. FlashFlow: A Secure Speed Test for Tor. In *2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS).* 381–391.

[32] The Federal Communications Commission (U.S.). 2021. The Federal Communications Commission (U.S.): Measuring Broadband America Program. https://bit.ly/3t34KZK. (Dec 2021). Accessed: 2022-04-22.

[33] The Federal Communications Commission (U.S.). 2021. Measuring Broadband Raw Data Releases - Fixed. https://www.fcc.gov/oet/mba/raw-data-releases. (Dec 2021). Accessed: 2022-04-22.