1. SCIENTIFIC TESTING: THE MOST RELIABLE WAY TO TEST A NETWORK

Executive Summary

The FCC is seeking to more closely regulate a key tactic in mobile carrier marketing—their performance and speed claims.

The commission already does this for fixed broadband and has proposed to use crowd data to set the upper limit for carrier marketing claims.

But here's the problem: There are significant differences between crowd and scientific testing.

Crowd testing is easier to conduct but tough to draw out any useful conclusions, while scientific testing takes significant resources to conduct but provides easy-to-understand and useful results based on a methodical process that is accurate and enables apples-to-apples comparisons. As a result, the FCC, in taking a shortcut with crowd testing, will not present the full or fair picture of the performance and speed of mobile providers.

Although the differences between crowd and scientific testing could just be chalked up merely to competition, with both sides advocating their approach, a major government agency has decided to throw its lot in with a crowd tester. Such an approach will provide a limited view of the mobile consumer experience and won't provide an accurate reading of the service providers' strengths and weaknesses.

In this report, we provide an overview of both scientific and crowd testing and provide a number of observations on the right policy direction.

The role of network testing and how it's changing

Network testing began as soon as the first wireless network was deployed. Initially, it was purely a technical endeavor. Engineers were curious to see how the network they worked to build was actually functioning compared with the computer model of the design. But because they were competitive in nature, the engineers decided to also test other providers—just to see how they measured up. The tests were based in science because scientists were the ones doing the testing—whether it was for their own network or a competitor.

At its core, network testing has a few important aims, including:

- Recording real-world conditions that operators' end-users experience
- Understanding network performance during the various activities of consumers
- Capturing an accurate representation of network conditions
- The ability to directly compare results from one operator to another

The early efforts were "scientific tests" that were an elaborate, difficult process that certain carriers continue to this day. It wasn't long before marketing departments seized on the testing as a vehicle to sell their services. At the same time, companies looked for a cheaper way to conduct tests without sending engineers out in a van for weeks at a time.

They were searching for something that would be "good enough" for marketers. The shortcut they arrived at is crowd testing, which relies on wireless users to self-select and do their own testing. Marketers see it as inexpensive and easy to implement. They can easily imagine positioning it as "real-time data from lots of 'actual' active consumers" across a wide geographic reach. If crowd data supports a marketers' desired claim, then they will use it, if not, then the data is ignored. Plus, crowd testing can potentially be scattered across a wide geography. Marketers either didn't realize or understand the limitations of crowd testing. Or maybe they just didn't care if it cut corners or didn't deliver the same set of scientific facts and figures as a scientific test.

Unfortunately, crowd testing doesn't include a number of factors that are critical to the consumer wireless experience. For example, crowd testing ignores the issue of reliability (i.e., crowd testing can't happen if the network is down or nonexistent). Moreover, it's limited to data, so it doesn't include a big part of the consumer experience—voice. Data is obviously important, but voice remains a critical part of everyone's wireless life.

To get to the heart of the issue, we investigated crowd tests and scientific tests. On the surface, they appear to do the same thing. But, as we have indicated, the methodology, along with the pictures they paint, are vastly different.

How scientific testing and crowd testing compare

Scientific testing utilizes a scientific process that replicates, with a standard device, a variety of conditions, and a vast coverage area all included. Scientific testing aims to go further than the boundary of the network—in fact, it can find the boundary (something crowd testing can't do). Moreover, when scientific testing is being conducted, all phones in the test make a call at the same time, under the same conditions, using the all the carriers being tested at the same time. Therefore, scientific testing creates comparable testing results that cover the people, where they travel and where they live. In fact, it even goes well beyond that. Crowd testing provides no such comparable data because it doesn't use a rigorous testing methodology that ensures the validity of the data.

With scientific testing, a vehicle or portable unit typically travels a tightly controlled, scientifically thought-out course around an area to establish a comprehensive picture (or statistical characterization) of the conditions. Every step is documented and the same tests are performed over and over again. The result is a comprehensive picture of the network (even where there is no connection) and the performance of the device on the network. By picking a standard device, the focus is on testing the network, not on testing devices. By focusing on capturing an accurate picture of the network—where the connections work and where they don't, scientific testing provides a better representation of the typical consumer experience.

On the other hand, crowd testing usually measures only a facet of that experience. And crowd testing often happens from a specific perspective or is based on special circumstances. Crowd tests rarely occur when a connection is average. Usually, someone initiates a crowd test when a connection is frustratingly slow or amazingly fast. One more thing: No one ever tests when there is no connection because the apps simply don't work in the absence of a signal.

Here's an interesting fact: A user who is wondering why a connection has slowed might find it is because they have exceeded their data allowance. The connection may, in fact, be slow. That's what happens when consumers reach the 3, 7, 10 or 20 GB limit for the month. But, interestingly, that slowness won't be reflected in all network crowd test data as a result of a speed test. Why is that? Because some carriers exempt the speed test from the slower speed. Such issues can be controlled in scientific testing. All useful and quality data is included for a full, accurate representation of performance, not just a selected view.

The lack of a methodology for crowd testing often results in a u-curve that distorts the typical experience into a bipolar world of great connections and lousy connection. This misrepresents the overall experience—either exaggerating the flaws or trumpeting the advantages. As a result, the *minority* of experiences are represented as the *average*.

The differences between crowd and scientific testing could just be chalked up merely to competition with both sides having to provide a rationale for their approach. That is, unless a major government agency decided to throw its lot in with a crowd tester—which is happening. The FCC, looking to more closely regulate how mobile carriers market their data performance and speed claims (something it has already done in fixed broadband), will use crowd data to set the upper limit for carrier marketing claims. As long as advertised carrier speeds (whether from their own crowd tests or more rigorous scientific tests) remain lower than the crowd data, the FCC won't object.

The FCC's data source of last resort is its own crowd-sourced data from an app that has received very little publicity or attention, which will translate into low usage rates of the app and very few samples. The upshot of the FCC's decision is that all the problems of crowd sourcing are compounded by a lack of a controlled, methodological approach. Inaccurate results from crowd testing could be highly magnified and over characterized. It's not about the sample size at all: If

the data is collected in a haphazard manner, whether it's a small sample size or a large sample size doesn't really matter. Nevertheless, it's important to note that the FCC's app was ranked 813th among Apple App Store Utilities, with other speed testing apps ranking 14th, 193rd, and 203rd among utility apps. The FCC plans to release results on all or most Cellular Market Areas[1] by the fall of 2016. The FCC's action is an endorsement of crowd testing, which is a test that is open to manipulation and many data quality concerns rather than being based on repeatable, scientifically crafted processes. It is also peculiar that the FCC would institute such a significant new policy without the appropriate review process that includes the opportunity to discuss the merits of a particular proposal. Instead, the FCC released the new policy without much fanfare as a mere guidance document.

Comparing the breadth and scope of scientific testing and crowd testing

Here's an interesting tree-falls-in-a-forest question for the crowd tester: If you have no signal can you have a test? The answer is, more often than not, *no*. The apps require a signal to even conduct a test. While a few tests record the failed test, many just don't bother with it. As a result, the failed test effectively never happened.

On the other hand, scientific testing happens in a controlled environment, using the same methodology every time—whether or not there's a signal. The result is a more accurate map that reflects real-world conditions and is free of the bias that is inherent with crowd testing.

Understanding coverage holes is as important as, if not more important than, knowing weak spots. If you are evaluating a carrier and don't see crowd data for your neighborhood, that might mean no one in your neighborhood has tested their phone. Or it might mean that there is no signal. It's unclear. It leaves the data open for interpretation.

That underscores a very important difference between crowd testing and scientific testing.

Scientific testing, with its proven methodology and fixed route, doesn't leave gaps. The testing, the methodology, the devices and the technology employed are the same everywhere.

On the flipside, operators generally do have coverage outside the scientific test footprint, but generally few people live there. If you are interested in rudimentary coverage and data-related performance, but not voice, then crowd testing may be your only option there.

[1] See definition: https://catalog.data.gov/dataset/cellular-market-areas

Mapping scientific testing and crowd testing

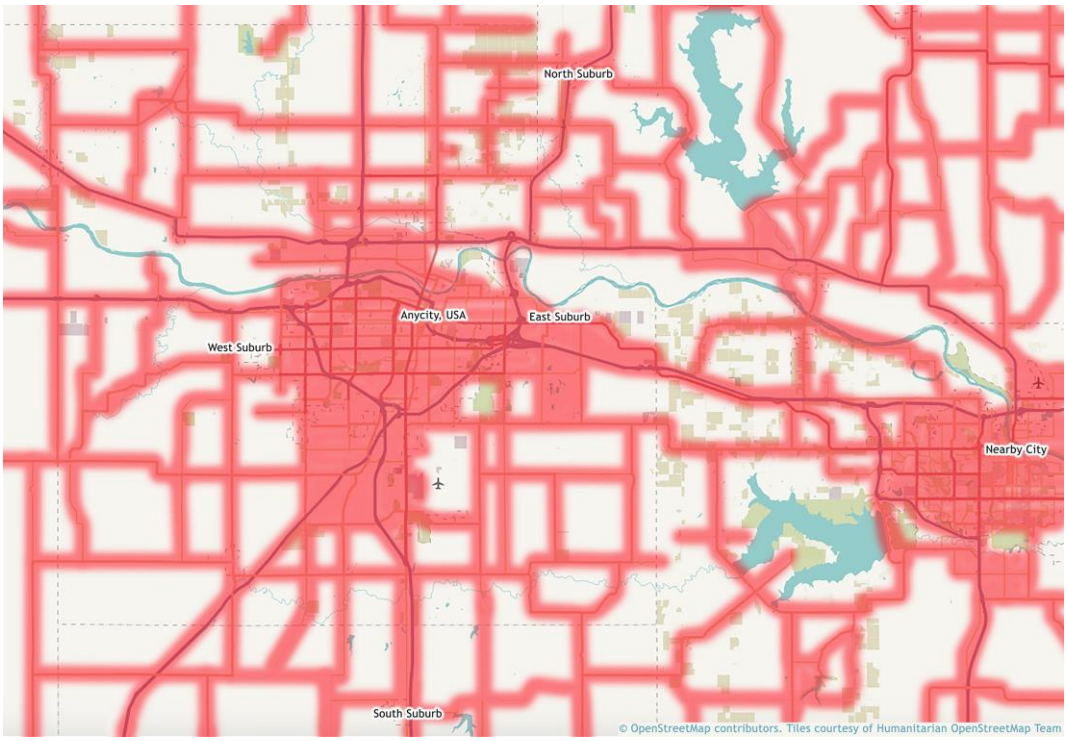
Let's think about the typical US city. We'll call it Anytown, USA. Like any city, it has a number of highways, a few suburbs, an airport and recreation areas. Even in a small city and its surrounding area, there are thousands of miles of road. A typical scientific test consists of somewhere between 1,000 and 3,000 miles in a city like Anytown, USA. Consumers rely on their phones and the networks on all of those highways, main streets and backroads.

Gauging the coverage a representative sample of those roads takes a methodical approach. With scientific testing, properly equipped vehicles use the kind of intelligence a delivery company puts into its routes to ensure each road and location is tested completely and efficiently in a controlled environment. The vehicles use a standard set of devices and duplicate real-world environments.

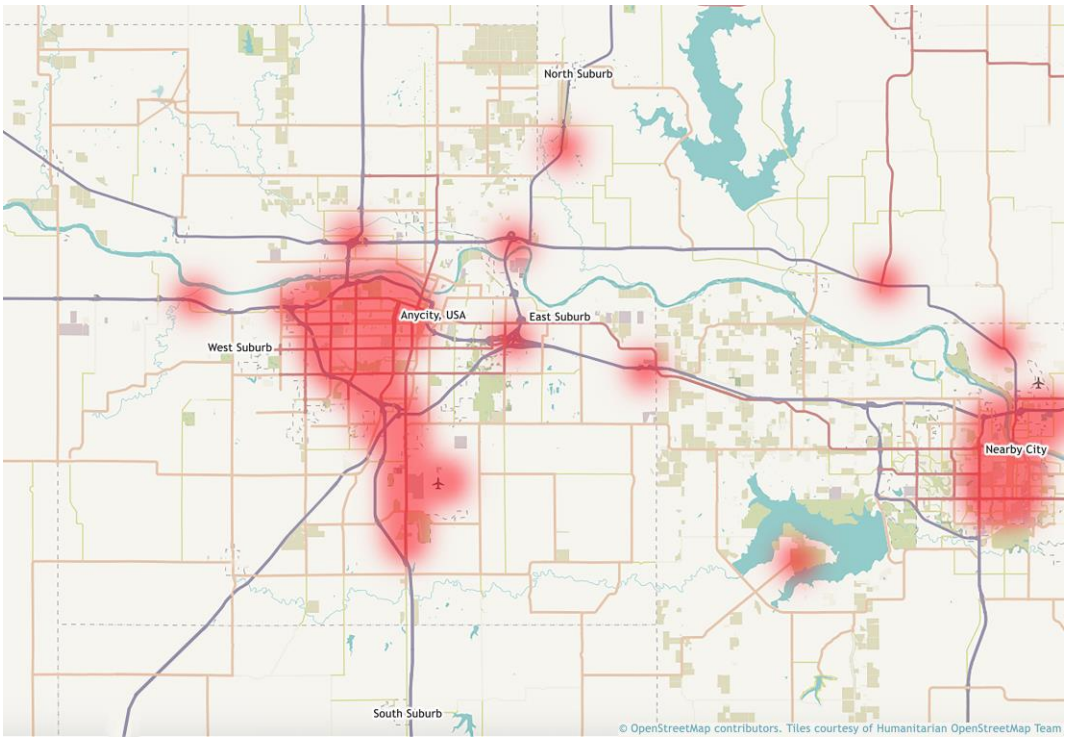
On the other hand, crowd testing employs a haphazard approach that has a bias toward technical people who tend to raise their hand and live in populated areas. The result is uneven representation of the coverage of a network.

In our Anytown example, the difference between scientific testing and crowd testing couldn't be more apparent (see Exhibit 1). Scientific testing covers a broad, representative area and crowd testing covers a few spots. It's fine if you live or travel in one of the spots covered by crowd testing. If you don't, you're out of luck. Even an increase in the crowd testing areas of 50% or 100% would result in a fraction of the area that scientific testing can cover.

Exhibit 1: The breadth of scientific testing…



…compared with crowd testing



What crowd testing leaves out

Crowd testing tells a one-dimensional story that is focused on data transmission rather than voice. While data is undoubtedly the growth segment of wireless consumption, it's hardly the only thing that matters. Despite all the prognostications that internet data and messaging is killing voice communications, nothing could be further from the truth. Rather than shift from one means of communications, we are just communicating more.

Today, Americans spend more than 900 minutes per month talking with each other, roughly the same amount as they did last year, the year before and the year before that. Messaging and internet data usage are increasing substantially, especially because the next generation of voice will use the data network as well. Voice over LTE (VoLTE) is a significantly improved version of plain old voice calls, but not everyone uses it yet.

Even with VoLTE on the rise, voice calls are still the dominant form of communication on most phones. Crowd testing completely ignores that and presents no information on voice quality and performance—not even something as simple as "Can you hear me now?" Of course, the crowd testing apps just wouldn't be good at this. Can you imagine how intrusive it would be to have an app making or receiving calls for you? Taking it a step further, it would be even more intrusive to have the app monitoring your calls for performance reasons. Simply put, it's too impractical and has massive privacy implications.

Scientific testing is built for this kind of thing.

Where does device bias come in?

The idea of network testing is, quite simply, to test the network. The best way to do this is by testing the networks with one standard device. One of the little secrets that consumers don't know—and nobody else in the industry wants to talk about—is that different phones connect differently to a network. It's like with people: Some have a firm handshake, some have a weak handshake, and some even have a slippery handshake. This can have a very significant impact on the customer experience, depending on the device the customer uses. But network testing does focus on the network not on the devices and the operator should get praised or blamed for the network they have built, not for the phones a device manufacturer has built.

Now, networks might be good or bad and phones might work well or not. But the reason crowd testing doesn't work reliably arises from the confusion over what the "connection" really is. It's not just one thing. And if you don't know what the connection is, then what are you testing? The device, the network, a combination of the two? Where does one end and the other begin?

The connection is a complex definition that lends itself to overly simplistic explanations. The device itself plays a big part in the network handshake. It can be the weak or strong link, just like the network. Which side of the handshake is weak is not a concern to the crowd test—it will always ascribe that flaw to the network.

The electronics differ in each phone—even those from the same device manufacturer. Some connect better and others have flaws that can slow down a connection before the network is even involved in a call, a text or a data transmission.

So, putting a nameless phone into the hands of a crowd tester on a confusing array of networks, with little if any control or established methodologies, standards or process in place, doesn't yield reliable results. And the device is absolved of any responsibility.

Self-selection and geographical/socioeconomic biases

Crowd testers self-select. They raise their hands and decide if, when and where they will test and report their phone. Often, the test happens because the connection is either very fast or very slow. Few people test when the connection is average.

In addition, results might be skewed to certain geographic areas. Crowd testing generally over represents major urban areas and overlooks sparsely populated areas.

As a result, it skews the data. Major urban areas have the typical bipolar results—with lots of highs and lots of lows, with very little in between. Suburban and rural areas are left with spotty crowd test results. Crowd testing focuses on more densely populated areas where younger people live, where people with more smartphones live, and it can de-emphasize where baby-boomers and older Americans live. Some carriers make coverage maps available and a quick analysis shows the selection bias—on roads where no one lives there are no crowd testing results.

On the other side of the ledger, true scientific testing blankets a region—regardless of geography, economics, or any other factors. The methodology and process used in one area—regardless of population, location or economic strata—is the same at every point in the test. The devices are the same, the testing process is the same, and the results compare apples to apples.

Identifying the right technology, location, and devices

The IP address is the central identifying detail for the connected world. Or so it would seem. It turns out that mapping IP addresses to specific carriers (and, therefore, validating that a particular connection was with a specific carrier) can be a significant challenge. Think of a carrier that has multiple lines of business—DSL, fiber, LTE. The IP blocks assigned to that carrier can produce a dizzying array of issues because many countries have converged carriers. If the crowd testing platform does not properly identify which IP address belongs to which carrier, the result can be a misattribution of the test result. This is less of a problem in the US, but can get quite difficult around the world when the delineations between wireless, DSL, cable modem, and fiber blur.

If a crowd test shows a 5Mb connection, it's often unclear if that is coming via LTE, fiber or DSL (or some combination thereof). A super-fast wireless connection reported by a crowd test might very well just be a very slow fiber or average DSL connection. That kind of doubt, which misrepresents the average performance of both the fiber service and the wireless service, rightly undermines the credibility of crowd testing and leaves the results open to interpretation.

Not only can a crowd testing app misrepresent geographies, it can often misrepresent or not even identify a specific location, which is critical for valid testing. Everyone knows there are coverage holes—some big and some small. That's just the way RF works. Move a few feet one way or another and the conditions can change. For instance, there are vast differences between indoor and outdoor performance. Crowd testing struggles to identify whether a test is outdoors or indoors, so it can't represent those holes accurately.

Even something as simple as identifying operators can be a challenge for crowd testing. One crowd tester had trouble properly identifying the operators consistently to enable an apples-to-apples comparison. For example, in Japan, the crowd testing provider correctly identified KDDI and Softbank on a nationwide basis. NTT DoCoMo, on the other hand, was identified by its nine operating companies instead of the nationwide brand. This made it impossible to properly compare the three carriers within the same area. In Germany, there was similar confusion about the reported speeds for Deutsche Telekom. Historically, the crowd service provider was able to distinguish between the results from T-Mobile (wireless), T-Home (residential landline customers), and T-Systems (business internet customers). Over time, the three sub-brands were merged into the Deutsche Telekom brand, merging the three vastly different technologies into one amorphous number.

The faults in crowd testing are evident in some recent data from another crowd testing provider that shows some vastly divergent results. First off, Virgin came in at 49.5% time spent on LTE and download speed of 4.05 Mbps. Boost came in at 57.2% time spent on LTE and download speed of 3.23 Mbps. Sprint came in at 50.9% time spent on LTE and download speed of 4.32 Mbps. That seems fair until you realize that the three carriers all share the same network—Sprints. Digging deeper into those figures, it becomes apparent that the LTE figure a device-driven statistic, not a network metric as the devices connect to the very same network.

It's too easy to game the system

In some cases, we saw the test results of pre-release devices, sometimes even from manufacturers that do not sell their devices commercially in the United States, on test networks with spectacular data speeds. Is it okay for someone to do a quick test or show off the pre-release devices on a test network to someone? Yes, of course it is. Is it okay to include such test

results as a record that should duplicate, or at least closely resemble the daily consumer experience? Of course not. Crowd tests are portrayed as real people doing tests with real devices. Including results from unreleased devices on test networks flies in the face of that.

In addition to the use of pre-release devices, anyone looking to game the system could create a script that conducts a test with a phone repeatedly, with each test appearing as a separate entry. VPN apps are freely available in app stores, so it's not a stretch to imagine someone using a VPN app to easily mask their location, then conducting tests don't reflect reality. It's not whether these approaches are possible. Of course they are. And even if only a handful end up in the crowd testing, they distort the results.

Looking around the globe, in country after country, we have seen the effects of manipulated crowd testing results. For some operators, for a certain time period, the number of tests was consistent, only to increase from one day to another by a fixed amount for a few days or even weeks (but only on weekdays). Then the number of tests would fall back to the previous level. In a specific example, the average number of tests was roughly 7,000 per day. That increased to 37,000 tests the next day, then dropped back to normal levels on the weekends, only to increase again on Monday, repeating the same pattern for three weeks and dropping back to historic levels for at least a year. Were the 30,000 additional tests valid? We don't know who was responsible for the sudden increase and if the mysterious increase was due to a test program for a specific device or an attempt to manipulate the results of the crowd testing. The 30,000 weekday tests should be discarded as non-valid, outlier tests because they quite clearly could have been done with an intent other than merely testing the network. Such events are concerning to say the least. So, the result with that crowd test is misleading at best.

On the other hand, scientific testing uses the same methodology every time in a controlled environment. It's critical to properly map coverage, as scientific testing does, or the picture is not complete.

Conclusion

Our analysis of difference between scientific testing and crowd testing brings us to the following conclusions:

- Scientific testing utilizes a methodology that is repeatable. It replicates, with a standard device, the typical consumer experience and covers a vast area.
- Every step of scientific testing is documented, with the same tests performed over and over again. Such a process ensures the validity of the test results.
- The difference in geographic coverage is stark: Scientific testing characterizes everything in a test area evenly, while crowd testing covers a few spots—some parts more, some less, and some not at all.
- Crowd testing usually only measures a facet of the consumer experience, often based on special circumstances or from a specific perspective. It rarely measures average, or typical, speeds.
- Crowd testing apps do not test for voice quality, or network reliability.
- Crowd testing blurs the line between the device and network—and ignores the complex handshake that takes place between the two.
- A connection that is slowed because a user has reached a monthly allotment will not reflect that because some network speed tests are not considered part of the data ceiling.
- The self-selected nature of crowd testing skews the data in such a way that there are mainly highs and lows and few typical results.
- The mobile ecosphere is complicated, with a staggering assortment of multiple national service providers and network providers, MVNOs, regional carriers and other players all in the mix. Scientific testing accounts for that with a documented methodology that employs a defined set of devices in a controlled environment that's managed by local market experts. On the other hand, crowd testing apps often have trouble distinguishing between network providers and service providers. Results for MVNOs that run the same brand on different networks can be convoluted while results for national providers that manage regional providers (and maybe an assortment of mobile, fiber, and DSL networks) can produce results that are far from reliable.
- When a scientific test is being conducted, all phones in the test make a call at the same time, under the same conditions, using the same carriers. As a result, scientific testing creates comparable testing results.
- Crowd testing doesn't provide comparable data because it doesn't use a rigorous testing methodology necessary to make such comparisons. In fact, crowd data is open to manipulation precisely because it is not comparable.
- In short, scientific testing is harder to conduct, but much easier to accurately analyze while providing easily comparable results. Conversely, crowd testing is easier to conduct but much harder to find any way to draw out accurate metrics representative of typical user experience, leading to anecdotal results that could be inaccurate and best and flat out wrong at worst.

This report was produced with financial support from RootMetrics.

ABOUT ROGER ENTNER

Roger Entner is the Founder and Lead Analyst of Recon Analytics.

He is known around the globe as one of the most respected telecom experts. Over the last decade he has been frequently quoted by the world's most prestigious media outlets, such as the Wall Street Journal, the New York Times, USA Today, Financial Times, ABC, CBS, NBC, Fox, CNBC, NPR, PBS, and CNN. In the last year alone, he was referenced more than 2,000 times. In addition, Roger's research has been cited in six Annual Mobile Wireless Competition Reports to Congress, making him one of the most quoted analysts in the history of these highly influential reports. His research around wireless spectrum has been cited by the White House's Council of Economic Advisers. Among his influential work over the last two decades, Roger has written four reports for CTIA documenting the increasing impact of the wireless industry on the US economy. Roger has a weekly video show with Wireless Week, and is a regular contributor to Fierce Wireless as well as commentator on RCR Wireless News' video shows, where he analyses and comments about customer and industry trends in the connected world. At age 45, Heriot-Watt University bestowed on him an Honorary Doctorate of Science for his contributions to the advancements in research of the telecommunications market making him one of the youngest Heriot-Watt University graduates to receive this honor.

Roger's main focus is the competitive telecom market place and how the market participants interact. He is one of the leading experts researching the wireless experience, how it influences customer behavior and how customers make choices.

Before starting Recon Analytics in January 2011, Roger was the Senior Vice President, Head of Research and Insights for the Telecom Practice of The Nielsen Company. With more than $5 billion in revenues, Nielsen is the largest consumer market research provider in the United States and around the world. Nielsen is also the largest market research provider to the telecommunications industry. In his role at Nielsen, Roger was responsible for advancing the research and thought leadership position of Nielsen in the world of telecommunications. In particular, he led the research regarding consumer behavior and consulted with the entire range of telecommunications companies—wireless operators, wireline telecommunications providers, cable television and internet service companies, mobile device providers and software providers—on how to improve their products and services.

Before that, Roger was Senior Vice President, Communications Sector at IAG Research and was part of the senior leadership team when Nielsen acquired IAG in April 2008. At IAG he was responsible for helping telecommunications providers improve the effectiveness of their advertising expenditures. Building on IAG's traditional strength in television advertising, Roger was involved in several successful engagements that expanded IAG's traditional television advertising effectiveness measurement to radio, the Internet, and mobile advertising.

Prior to joining IAG Research in 2007, Roger launched the North American coverage for Ovum as Vice President, Telecom. He established the company as one of the leading telecom research

providers in North America. Before joining Ovum, Roger headed the wireless carrier research group at the Yankee Group from 2001 to 2004.

From 2002 to 2003, he was a member of a 16-person SBIR/STTR Phase II Panel for the National Science Foundation. He helped direct federal research grants to innovative, high-risk projects with a significant potential for commercial viability.

At both the Yankee Group and Ovum, Roger focused on researching trends in the wireless world and advising clients on current and emerging business and consumer trends that affect the wireless world.

Previously, Roger was Strategic Marketing Manager for LCC International, which designed and built wireless networks around the world. In that role, he assessed the trends and developments in the wireless world and developed strategies to help LCC benefit from emerging opportunities. Part of his focus was understanding and determining the demand for cell sites based on coverage and capacity requirements based on customer behavior.

Roger received a Bachelor of Arts in Business Organization, from the Heriot-Watt University in Edinburgh, United Kingdom, a Master of Business Administration from the George Washington University in Washington, DC, and an Honorary Doctorate of Science from Heriot-Watt University.