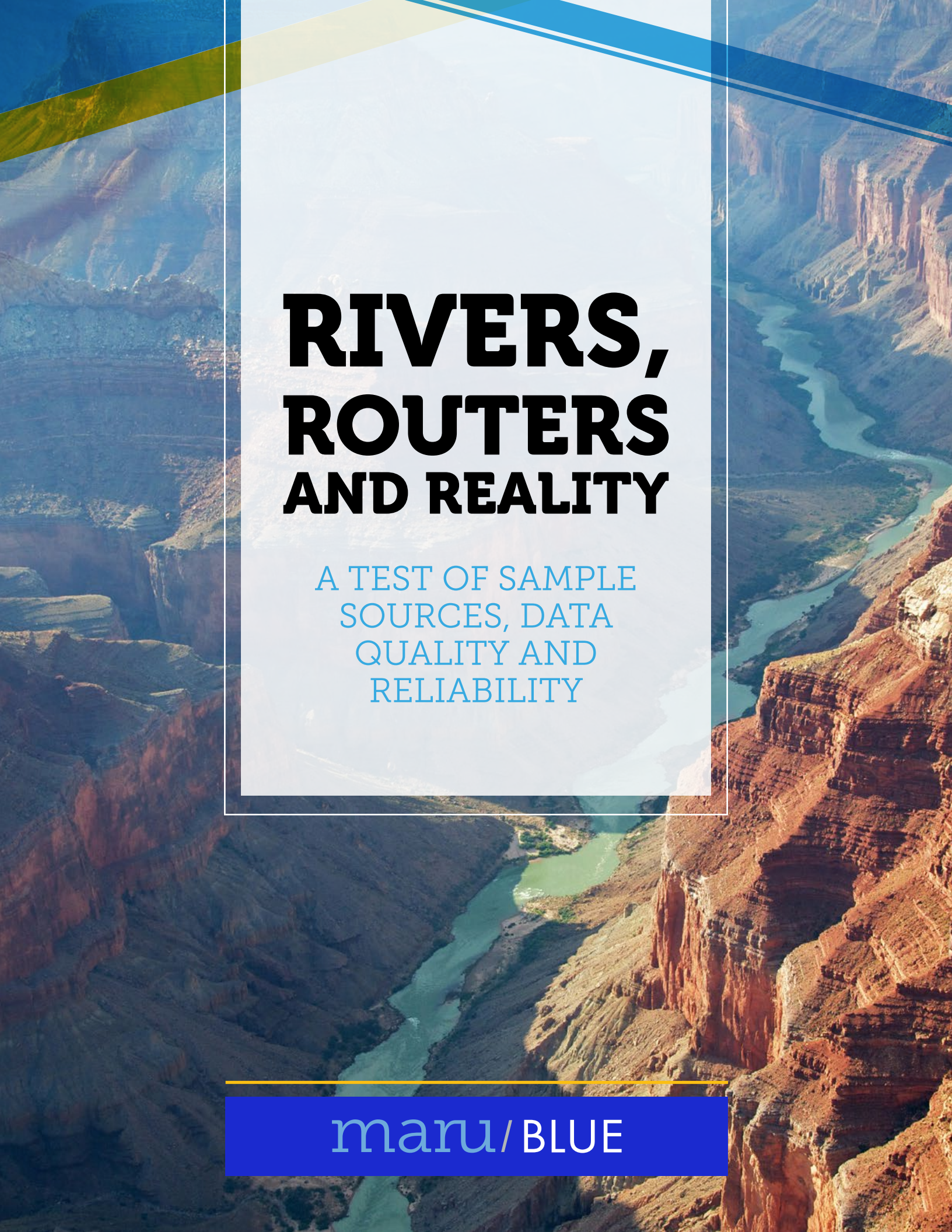


RIVERS, ROUTERS AND REALITY

A TEST OF SAMPLE SOURCES, DATA QUALITY AND RELIABILITY

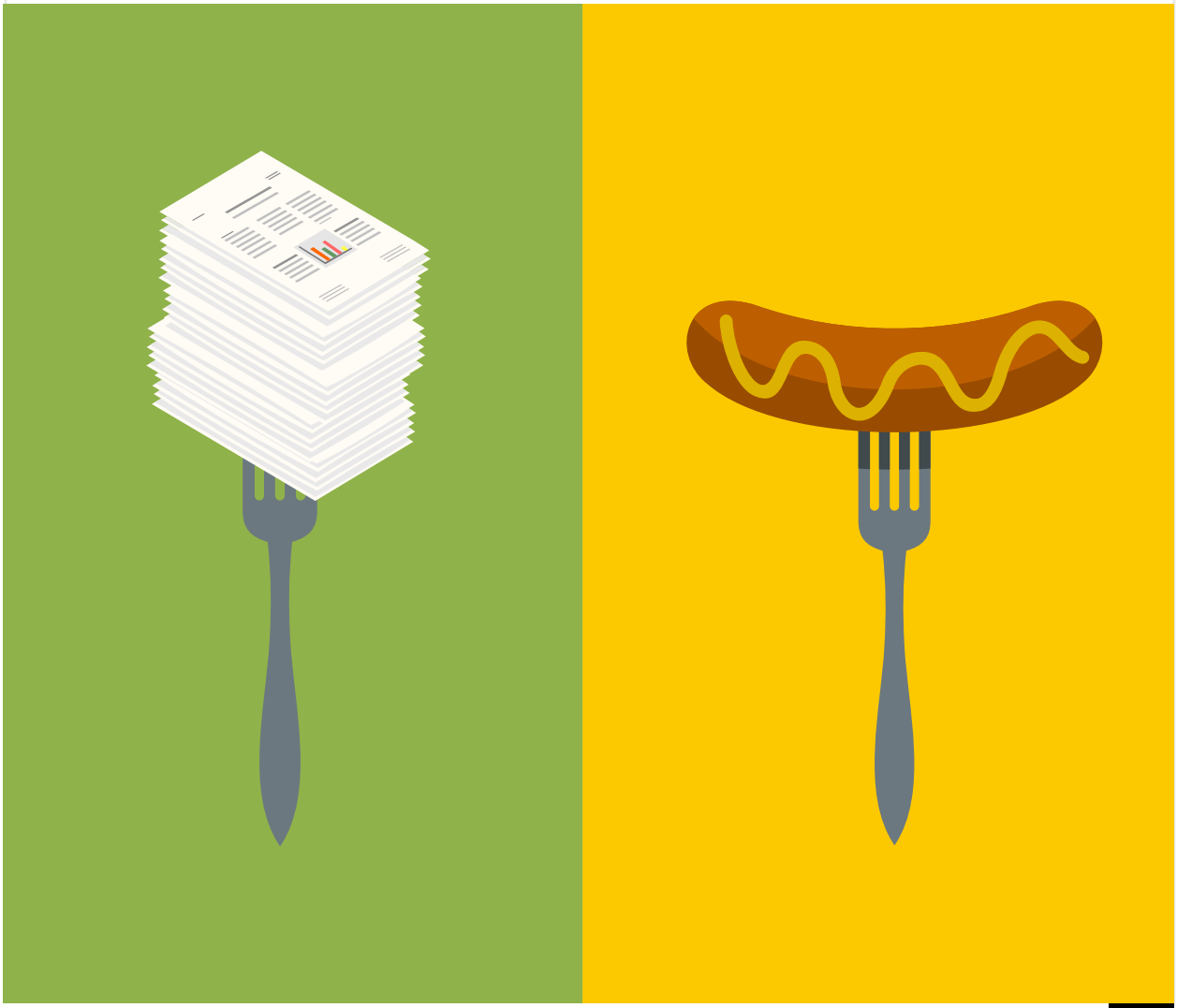
maru/BLUE

SAMPLE AND SAUSAGES HAVE MUCH IN COMMON

Our research insights are only as good as the sample upon which the research is based. But sample is often treated as a commodity. Should we take online sample for granted?

It's an important question to ask and answer, because sample quality determines the reliability and reproducibility of our research. If our sample is bad, the conclusions we draw from our research could harm rather than help.

In this whitepaper, we report on research into the quality of online sample, drawing on numerous studies. The results make it clear that all sample is not equal and that those making decisions based on research should ask some critical questions before they put their trust in the data.



Our investigation produced five key findings:

1. Sample is often sold and resold, and the company you are buying from is frequently not the original source.
2. Asking the same questions of a demographically representative general population sample can yield very different results, depending on the sample source.
3. The source from which a panel is recruited can profoundly skew the results and make the data unrepresentative.
4. Many sample sources do not deliver reproducible results—if you ask the same question twice you get differing results.
5. Routing sample is a new and increasingly common practice, but there is data to suggest it can reduce the reliability of the results.

These provocative findings lead us to call out "caveat emptor"—let the buyer be aware.

We recommend that those relying on online sample for research insights ask these three questions about every sample source:

1. What sources was/is the panel recruited from?
2. Is routing being used? What kind of routing is it?
3. Is there evidence of the reliability and validity of the sample?

But don't just take our word for it. Read on and see the evidence.

IS ALL SAMPLE THE SAME?

In statistics, survey sampling describes the process of selecting a sample of elements from a target population to conduct a survey. In everyday practice, online sample is often treated as a commodity, with price and ease of availability driving choice. But is that a good idea? Is all online sample the same?

In this whitepaper we report on some recent research on sample quality, reliability and reproducibility, drawing on studies by Maru/Matchbox and the Advertising Research Foundation's recent Foundations of Quality 2 initiative. We discuss the implications for the research industry and leave readers with some important questions to ask when making choices about sample.

Where does your sample really come from?

The world of online sample has a bewildering myriad of players. There are agencies, suppliers, sample validators, brokers, and trade desks. They buy, sell, swap, validate, resell and repackage online sample. But does the name on the package, match the contents? If you are buying sample from company X, is it really from them or is it coming from somewhere else?

Sample and sausages: how are they made?

"Those that respect the law and love sausage should watch neither being made," wrote Mark Twain. One might be tempted to say the same about sample. It's complicated and confusing and maybe sometimes you just don't want to know exactly how you got it.

But we can't afford to do that when sample quality determines the reliability and reproducibility of our research.

If our sample is bad, the conclusions we draw from our research will mislead those who need our insights.

WHEN THINGS GET DESPERATE

"...when she needs to find the last few respondents to finish the study, she will turn a blind eye and tell the sample provider to just finish the job and find them. "[...] hide my eyes, close my mouth -- get me those 50 people, and I don't care how you do it."

Excerpt from the Advertising Research Foundation's (ARF) Report on Focus Group Findings ARF FoQ 2 Router Initiative, 2014 by S. Gittelman and E. Ribero

WHEN HORSEMEAT BECAME BEEF: A CAUTIONARY TALE

In recent years there was a scandal when it was discovered that what was being sold as beef was actually horsemeat. "Horsegate" involved some of the biggest brands in Europe. "More than a dozen nations have detected horse flesh in processed products such as factory-made burger patties, lasagnas, meat pies and meat-filled pastas. The investigations have been complicated by elaborate supply chains involving cross-border middlemen" reported CBC News.[i]

Investigations revealed that correctly-labelled horsemeat was sold from two Romanian abattoirs to a Dutch meat trader, who sold it on to a Cypriot company that, in turn, sold it to a French firm who sold it on to a company in Luxembourg. Somewhere along the way, the horsemeat was relabeled "beef."

When it finally reached food manufacturers, it ended up in foods sold by such industry giants as Nestle, Ikea, Tesco and Asda.[ii]

Insert "sample" for "processed food products" in the following report on the horsemeat scandal and see if it sounds familiar:

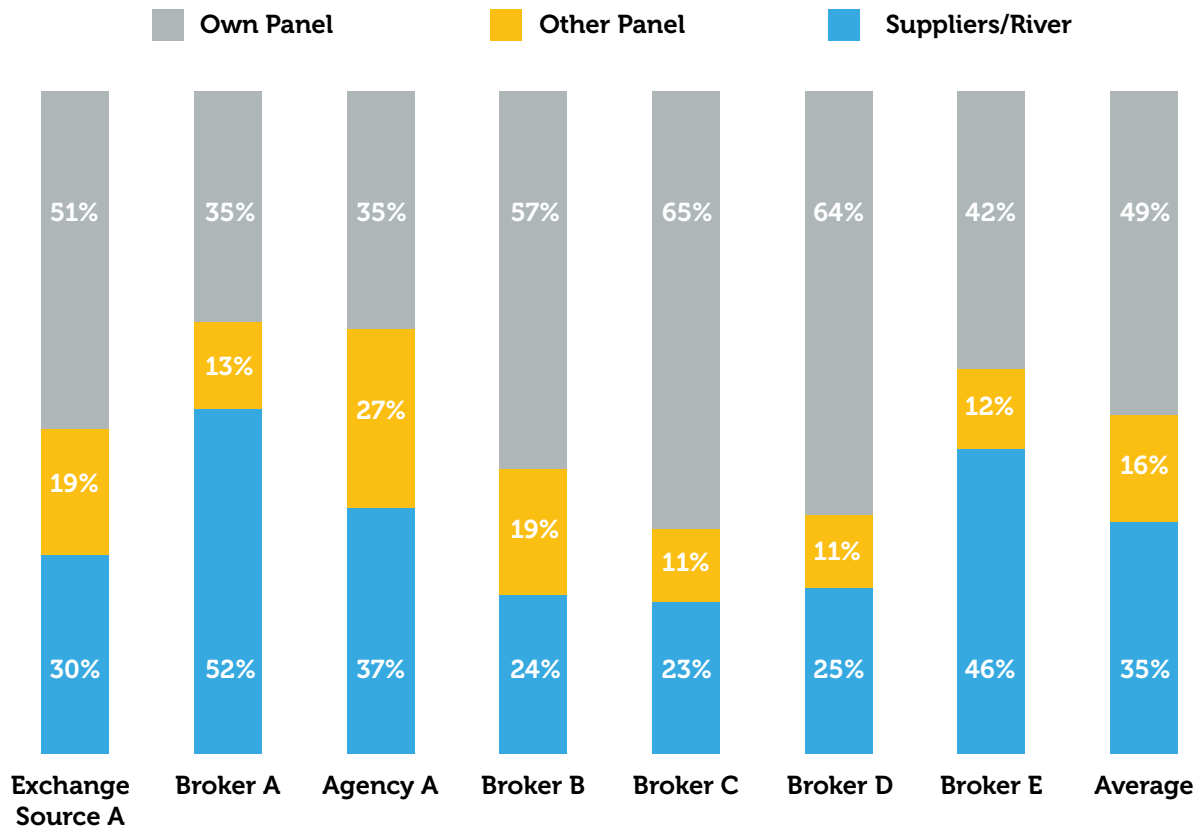
"Processed food products — a business segment with traditionally low margins that often leads producers to hunt for the cheapest suppliers — often contain ingredients from multiple suppliers in different countries, who themselves at times subcontract production to others, making it hard to monitor every link in the production chain."[iii]

DO YOU KNOW WHERE YOUR SAMPLE COMES FROM?

With the multitude of interconnected sample providers, brokers and trade desks, the actual source of a sample is often murky. In fact, there is evidence to suggest the source is often not what it appears.

Millward Brown Digital conducted an analysis that traced the original sources of sample purchased from seven of America's leading agencies, brokers and exchanges. They found that, on average, only one third of the sample they purchased was from the original source (e.g. "Agency A") and that the remainder came from either the River, another panel, or some other place. Do you know where your sample is from?

Source of Sample Sold



SURVEYS AS SCIENCE

Many argue that, in practice, sample source doesn't matter all that much, that it is basically interchangeable and that any reasonable source is acceptable—when used in a consistent manner. But is that assumption correct?

Survey research is an offshoot of the social sciences. Among the pillars of science are reliability and reproducibility. In the world of survey research that means you should be able to conduct a study with different sample sources and come to the same conclusion (reliability) and, that if you repeated the study you would be able to get the same results (reproducibility).

Knowing that the value of all our industry's insights depends upon the reliability and reproducibility of our methods, we set out to test various North American sample sources for reproducibility and reliability.

For more information on these studies, see the note at the end of our white-paper titled "How did we do this?"

Rivers, routers and repeatability

This inquiry is not exactly a unique endeavor. Major industry groups, have looked at panel sample quality before. Dutch group NOVPO's produced pioneering work in 2006 and there were seminal studies by Advertising Research Foundation in 2008 and Marketing Research and Intelligence Association in 2009.

But 2009 is a lifetime ago in the nascent world of on-line sample. Even the Advertising Research Foundation's Foundation of Quality 2 (FoQ 2) study—just being released at the time of this writing in 2014—was fielded in 2012.

Since then, River sample has become increasingly ubiquitous and routing has become part of the everyday reality of the sample world. Times have changed.

WHAT IS RIVER SAMPLE?

The name River sample evokes visions of pristine waters flowing softly, with babbling rapids and a meandering path through a verdant forest. But would you drink that water straight from the mountain river without knowing where it came from?

River sampling is an online sampling method that drives potential respondents sourced from panels or ads or pop-ups on social media and other websites to a router (see next page), where they are screened for studies in real-time.Qualified respondents are then assigned to an appropriate survey.Because of the screening you can ensure the respondents match your demographic requirements.

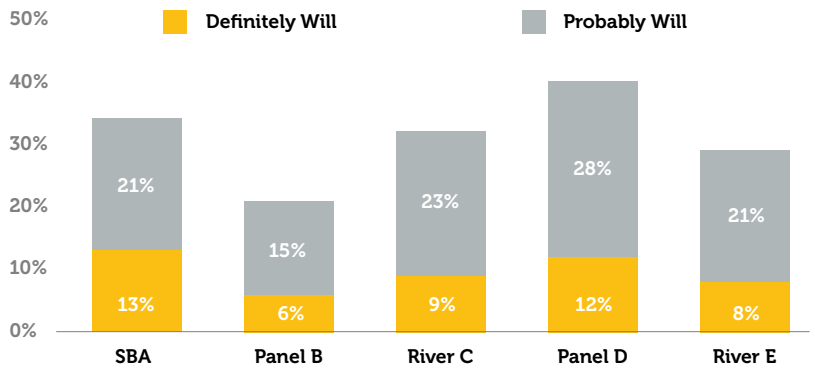
RELIABILITY - ARE RESULTS SIMILAR?

Our first study compared our Springboard America Market Community with two sources of River sample and two other panels. The sample for this study was nationally representative and the questions were quite general, including a few items about snacks and exposure to a concept for a new type of snack.

There were differences between the samples on a number of measures, but the variability in this chart provide a taste of how inconsistent the results were. This kind of discrepancy can cause some very practical problems. If this were your new product, would you recommend it go forward, based on the 40% likely to purchase from Panel D?

Or would you scrap it because of the 21% score from Panel B?

Study 1: Likelihood to Purchase Snack



WHAT IS A ROUTER?

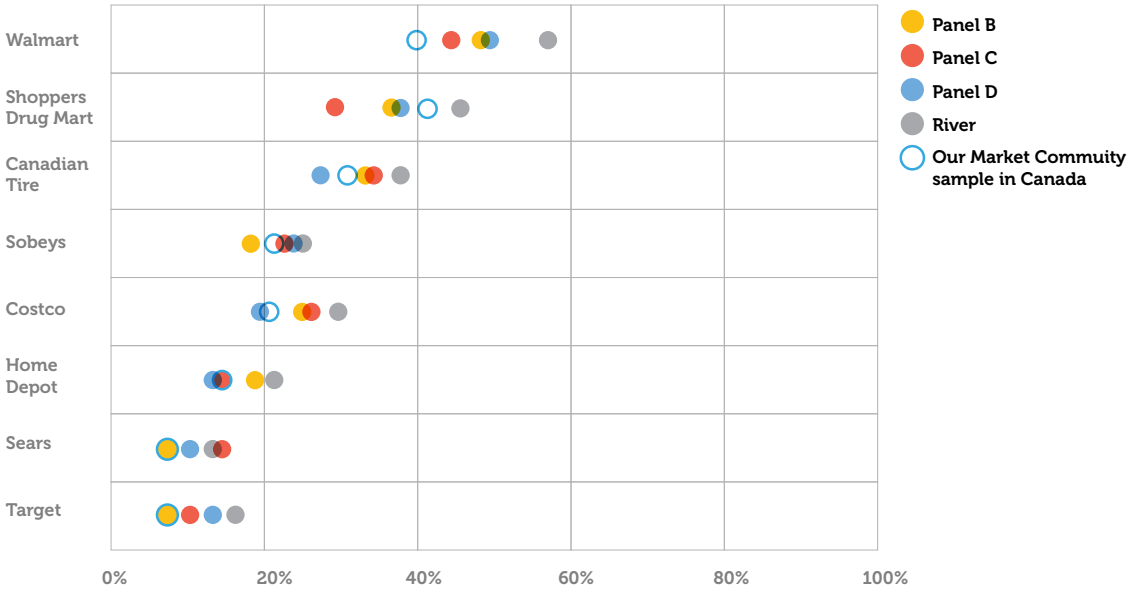
"A technology-based mechanism for allocating online surveys placed on the router to a sample of respondents who enter the routing environment.

Routers use software and business rules…to assign online surveys to a stream of incoming potential online respondents who enter the environment, usually at a general "landing page"…. In some environments, as soon as a respondent is found to meet all of the qualifications of a particular study's screening criteria, the respondent is passed onto that survey. In other routing environments, respondents may complete the criteria of several screeners and a process selects a survey for the respondent from among the ones he or she is found qualified to take."
Source: ARF FoQ 2

A second study, also with a general public sample, but this time from Canada—gives us additional insight into how variable results can be based on sample source. Here we compare 4 panels—including Maru/Matchbox's Canadian Community— and a River sample.

The questions in this second study were also quite general, with a number of them looking at mass merchandisers in the Canadian market.

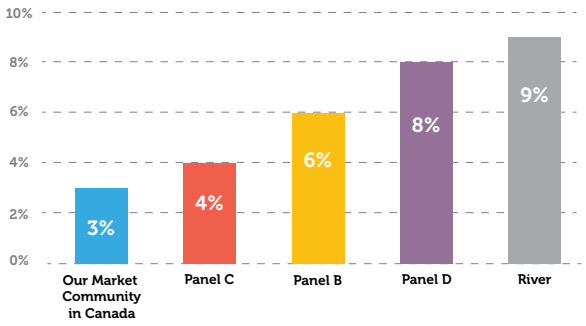
Study 2: Shopped at Store in the Past Month



We see tremendous disparity again, particularly on the more commonly shopped stores. And the differences are material. Did 25% of people go to Shoppers Drug Mart last month, or was it more like 45%? If that was your store, or your competitor's store, would that kind of variation be troubling?

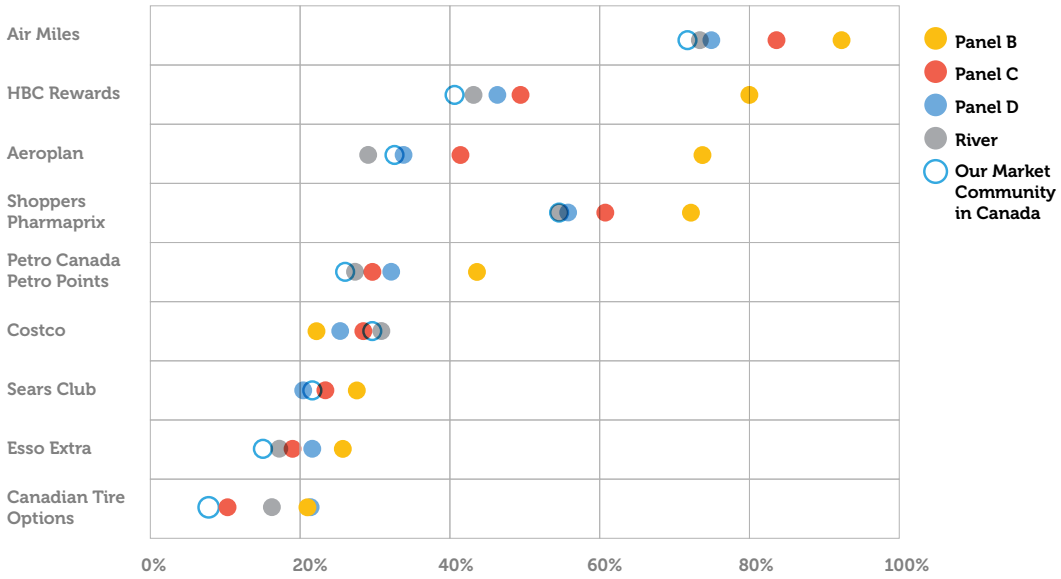
The higher levels of claimed use on the River sample might be partly explained by a higher rate of straight-lining we saw across the questionnaire with the River sample.

% of sample Straight-lining



In the second study we also looked at rewards card usage. We saw such notable differences that it would appear that Panel B may source some of its members from reward card companies.

Study 2: Ownership of Rewards Cards



If Panel B is so skewed on rewards cards, what implications might that have on studies testing promotions or rewards or couponing or anything incentive related?

And who knows what other biases might be linked to belonging to particular rewards programs.

REPRODUCIBILITY - DO WE GET THE SAME RESULTS TWICE?

Being able to reproduce results is what separates real science from snake oil. The history of science is littered with results that could never be replicated. One of the more famous was "cold fusion," a type of nuclear reaction that was claimed occurred at, or near, room temperature, compared with temperatures in the millions of degrees that are required for "hot" fusion. A couple of scientists--Martin Fleischmann (then one of the world's leading electrochemists) and Stanley Pons—reported that they had been able to achieve cold fusion. In the press conference reporting their extraordinary findings Fleishmann and Pons, backed by the strength of their scientific credentials, repeatedly assured the journalists that cold fusion would solve environmental problems and would provide a limitless inexhaustible source of clean energy, using only seawater as fuel. They said the results had been confirmed dozens of times and they had no doubts about them. But what might have spurred a revolution in generating energy turned out to be just a mistake, when dozens of scientists from multiple labs tried but failed to replicate the results of the original "cold fusion" experiment.

So what effect does sample source have on our ability to reproduce results? Are variations in reproducibility leading some researchers to report "cold fusion" like results in their studies?

To see how reproducible different sample sources were, we conducted two experiments. One was a replication of the second study reported on above. The other was an analysis of a very large tracking study.

We wanted some large sample sizes and multiple waves close together, so for our third study we looked at the results of a tracking study that used dozens of sample sources and had over 19,000 people per wave.

When we look at reproducibility, we find that over half of the sample sources show unexplainable, yet statistically significant, differences for each wave.

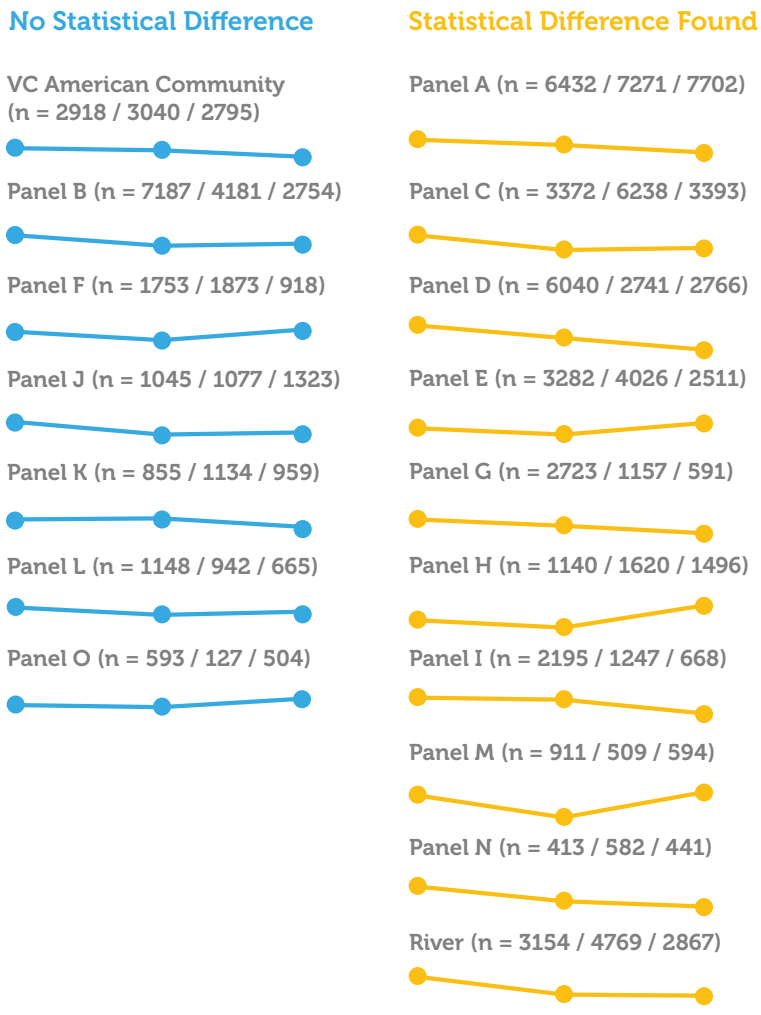
Given the large sample sizes involved, these results certainly suggest that sample source can have big implications for reliability.

In a fourth study we replicated the results of the second study—the one run in Canada with 4 panels and a River sample. The study was originally fielded in May 2013 and then replicated with the same questionnaire and sample sources in June 2014.

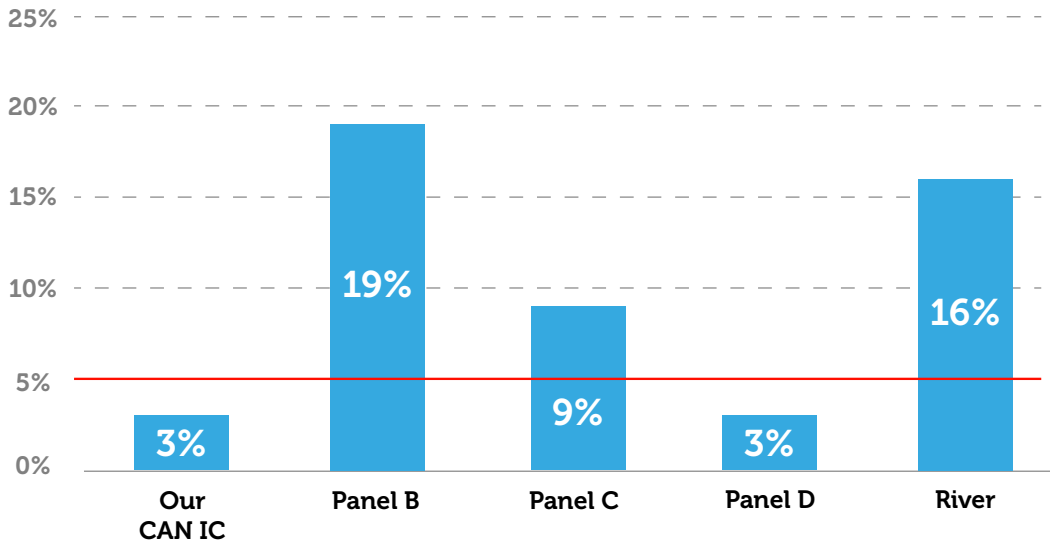
We found that, for two of the sample sources, we saw statistically significant differences on just 3% of the questions— within the 5% margin of error we would expect. But for three of the sample sources we discovered larger variations from year to year— two to four times more differences than we would expect by chance.

Study 3: Comparing the Reproducibility of Tracking Results

Average Familiarity Across Brands
Wave 1, 2, and 3



Study 4: What % of answers from study 2 differ significantly when we replicated the study one year later?



The low level of reproducibility we see with Panel B, the River sample and, to a lesser extent, Panel C would certainly put you at risk of reporting "cold fusion" type findings—results that simply can't be replicated.

What are we providing?

This lack of reproducibility with some sample sources poses a significant problem for our industry. Not only does it risk making tracking studies a mockery, it puts the reputation of the insight function at risk.

If our results can't be reproduced, how can the insights and recommendations based on them be valid? Are we providing scientific evidence or just selling snake oil?

ROUTING AND THE RIVER

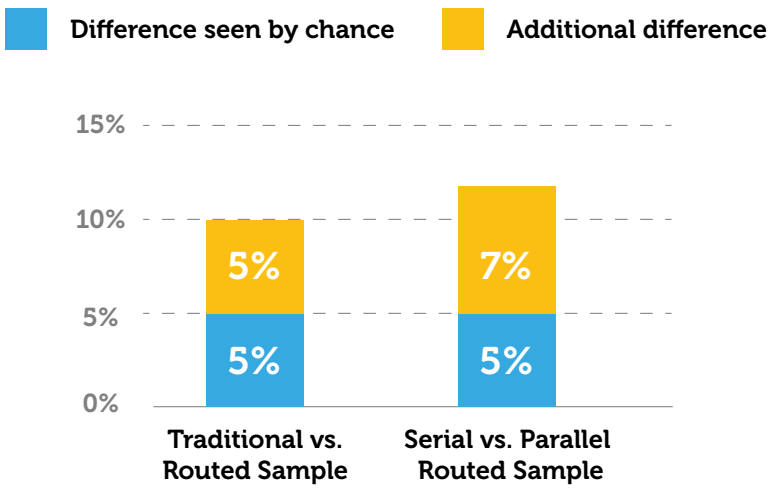
One of the aspects of today's sample marketplace that plays a role in sample quality is routing. It is widely used but not necessarily done the same way everywhere. In a recent review of the literature, the authors of Online Panel Research: a data quality perspective note "...there is a good deal of variation in how these router systems are designed, how they operate and what impacts, if any, they have on the data."[iv]

This variation makes it hard to generalize about the effects of routing. It is probable that the effects will vary from system to system and that, even when using a common piece of software, the impact will differ based on the rules the sample supplier applies when using the software. Nonetheless it is worth looking at the results of some research on the effects of routing by the Advertising Research Foundation (ARF).

In a recent Foundations of Quality 2 report they describe the results of a pair of important tests. In the first one they compared sample that had been routed with "traditional" sample that was pulled and deployed. In the second test, they compared serial vs parallel routing.[v]

The results provide cause for concern. In the comparison between traditional sample and routed sample, the researchers saw statistically significant differences on 10% of questions—double what you'd expect by chance.

ARF FoQ 2 Router Initiative Test Results



In the comparison of serial routing vs parallel routing, differences were detected on a total 12% of questions.

These results are not encouraging, but they do suggest that some of the lack of reliability and reproducibility we are seeing are driven by routing, as well as by other factors.

The authors of Online Panel Research note that, despite variability in approach and a paucity of evidence of reliability that "in all probability, [routing] will become the dominant paradigm in the next few years. The impact, if any, on data quality is unknown." That last line is somewhat chilling, especially when you consider how little most researchers know about routing, how it is done and what its effects are.

As part of the ARF FoQ 2 Router Initiative a focus group of researchers was conducted. They found that when talking about routers:

> *[The Researchers]* **"want to know more about how the router technology itself works, including the sample sources being fed into them. The problem is that they don't know exactly what questions to ask.**
>
> **Without knowing what else was happening within the router (other studies, volume, etc.) there was a sentiment that they were unable to tackle the task themselves. One participant suggested that this was the sample provider's responsibility, while others thought that the panel companies were not researchers and wouldn't know if there was a bias being created in a survey by using a router.**
>
> **One of the participants also" explained that the problem she has with the routing providers is that they are not researchers and their main priority is profit, stating "What they're doing there -- their bottom line is to make money based on selling the completes."**

That's a bit scary, is it not? The researchers don't know what's going on, there is variation from company to company, and even from study to study, in terms of how routing is done and what rules are being applied.

With the variation in panel quality that we see, we already are chipping away at the scientific foundation of our industry. If our information cannot be reliably used to make business decisions, it is helpful or harmful? What value are we bringing to the business?

With routers, blindly embracing an often idiosyncratic technology without careful research seems very risky—especially since the main reason for doing it is to make already inexpensive sample just a few cents cheaper.

LIVING IN THE REAL WORLD

We all know that online sample is imperfect, yet somehow the industry soldiers on. But are we headed in the right direction? Are we helping or hurting the insight cause when we use poor quality sample?

We all have our choices to make. And we all decide what we want to invest in making sure we are producing reliable and reproducible data.

But at the very least, we feel compelled to call out "caveat emptor." Let the buyer be aware, so they can act accordingly. Unless there are demonstrable proof of the reliability and reproducibility of a sample source, the results should be regarded with suspicion—like "cold fusion." And unless your sample supplier is open and clear about the actual source of your sample, you should wonder if the "beef" you are buying contains horsemeat.

Here are three questions worth asking about every sample source:

1. What sources was/is the panel recruited from?
2. Is routing being used? What kind of routing is it?
3. Is there evidence of the reliability and validity of the sample?

The sample suppliers as an industry have done a good job meeting the many challenges in the era of online research. The many innovations that have come aboard (panels, rivers and routers) provide the very soil market researchers need to grow their projects. But as we embrace change, let's be very conscious of the fact that the insights we derive from our research are only valuable when they are reliable and reproducible.

We hope this review of evidence on panel sample quality helps us all be more aware of the good, the bad and the ugly of sample quality. Caveat emptor!

TO LEARN MORE, CONTACT:
sales@marublue.net

ENDNOTES

i Horsemeat found in Ikea meatballs in Europe. The Associated Press. Feb 2013
ii 2013 horse meat scandal, Wikipedia
iii See endnote I
iv Online Panel Research: a data quality perspective M. Callegaro, R. Baker, J. Bethlehem, A. Goritz, J. Krosnick and P. Lavrakas Eds. Wiley 2014
v "In its most basic form, parallel routing is a process where a respondent is exposed to a set of pre-screening questions from all or a subset of the surveys in the routing environment, simultaneously, on the same page. After the respondent answers these pre-screening questions, he or she is assigned to one of the surveys for which the person appears to be qualified. At that point, more specific screening may be conducted before the respondent is deemed eligible to complete the survey.

In its most basic form, serial routing is a process whereby a respondent is screened sequentially for the available studies in the routing environment. Upon qualification for a study, the respondent is often immediately routed to that particular survey. Upon disqualification during the screening process, the respondent is reallocated, that is, he or she is re-routed to another screener (associated with another study within the environment). The process repeats until the respondent meets the qualification for an active survey within the routing environment, OR, based on business rules (e.g. length of time in the process, number of screening questions respondent has answered by clicking, number of screeners respondent has been administered, the process ceases and the respondent is thanked and terminated." Source ARF FoQ 2 Router Initiative Glossary

How did we do this?

All interviews were conducted online with adults age 18+. We had quotas for age, gender and region that were the same for all sample sources. When comparing across sample sources, all sources saw the exact same questionnaire on the same survey interface. The data were weighted, within each source, to ensure they were representative in terms of age, gender and region. A confidence interval of 95% was used for all tests.

Study 1 was conducted in the US November 3-9, 2011 with a sample of 2514, with roughly 500 each coming from Springboard America, three other panels and a River sample source.

Study 2 was conducted in Canada between April 30, and May 13, 2013 with a sample of 1,580 with roughly 300 coming from our Canadian Market Community and, three other panels and a River sample source.

Study 3 consisted on 3 waves of a brand image tracking study conducted in the US. Wave 1 had a sample of 19,229 and was collected Jan 14-20, 2013. Wave 2 had a sample of 19,222 and was collected February 18-24. The third wave had a total sample of 19,605 and was collected April 29-May 5, 2013.

Study 4 was a replication of study 2, and conducted in Canada in May/June 2014 with a sample of 1,580 with roughly 300 coming from our Canadian Market Community and the same three other panels and River sample source used in Study 2.