WHITEPAPER

# Data collection techniques for quality outcomes

*New research by Rep Data, Research Defender and DM2 uncovers best practices for primary researchers*

There's no arguing the point that making data-driven decisions is critical to organizational success. In fact, a recent article in Entrepreneur Magazine indicated that businesses that use data are three times more likely to "significantly improve their decision-making process than organizations that do not." Consumer insights, one vital part of a company's data landscape, can provide clarity of purpose, and confidence when executing business-critical strategies. But those insights must be firmly grounded in a market research process that is, in turn, firmly grounded in techniques that deliver quality data.

The unflagging conversation surrounding quality data in the market research industry generally centers around a few key concepts executed during the data collection and fieldwork stage.

**With this project, we set out to test the impact of three specific techniques on data quality: sample sourcing expertise, fraud mitigation technology, and research project management.**

## SAMPLE SOURCING

Unbiased and representative sample, garnered from diverse sources, is the basic foundation for quality outcomes. When it comes to a diverse sample, approaches run the gamut. Some market research projects draw on single proprietary panels, with little to no sourcing of sample from outside sources. At the other end of the spectrum are large exchanges that integrate real-time sample with private panels and everything in between, using automated techniques.

Our hypothesis for this research project is that the sweet spot for this piece of the puzzle lies in the middle, coupling hands-on, expert project management with targeted, representative sample identified from a wide variety of sources. To explore this further, Rep Data sourced respondents from four sample providers, indicated in this research-on-research as Providers A through D, to provide a basis for showing the impact of intelligent sample sourcing on quality outcomes.

## FRAUD MITIGATION

The next step in achieving quality is putting the expertly curated sample through a gauntlet of fraud mitigation and respondent engagement techniques to further ensure positive outcomes. There are a wide number of approaches in the market research industry today designed to mitigate fraud. From traditional, tried-and-true methods to newer AI-driven algorithms, blending techniques for ensuring data quality requires a deft, balanced approach. As digitalization accelerates, fraud instigators have more avenues for their untruthful activities, using the tech at their disposal to "trick" the system.

We tested out a number of techniques, offered by Research Defender, during this study, applying them individually and in unison to determine impact on quality measures.

- **Digital Fingerprinting and Fraud Identification:** Called "Search" in this paper, this proprietary technique examines potential respondents based on their past external activity before they engage in a survey.

- **Text Analytics:** Another proprietary solution, called "Review," measures and scores a respondent's engagement in real-time by analyzing overall quality and thoughtfulness of open-end responses.

- **Respondent Level Tracking:** Called "Activity" here, this method tracks a respondent's rate of activity across the market research ecosystem, and is useful in flagging professional survey takers.

These techniques were further enhanced by sample supplier Rep Data's stringent quality control standards, applied to each panel provider included in the study, and furthered by the company's standard automated and personal quality checks at all points along the project lifecycle. This level of personal monitoring of sample and study quality through the whole process allows continuous improvement, relying on Rep Data's research project management expertise.

### KEY TAKEAWAYS

- Layering fraud mitigation techniques positively impacts outcomes by creating a clean, healthy and efficient market research ecosystem
- Unbiased, efficient sourcing from multiple panels and sample suppliers delivers more representative results
- Using expert project management for fieldwork eliminates common challenges in the data collection process

## STUDY METHODOLOGY

This project was conducted by research veterans at DM2 to assess the efficacy of applying different screening and data quality techniques in a survey setting.

A 13-minute online survey was conducted in April 2021 among n=2,002. Rep Data sourced equal sample from four of the industry's larger online sample providers. Completes were evenly distributed across five cells, with providers delivering n=100 to each cell with consistent age and gender quotas.

| Provider | Raw Data | Search | Activity | Review | Layered Approach |
|---|---|---|---|---|---|
| A | 100 | 100 | 100 | 100 | 101 |
| B | 100 | 100 | 100 | 100 | 100 |
| C | 100 | 100 | 100 | 100 | 100 |
| D | 100 | 100 | 100 | 100 | 101 |
| **Total** | **N 400** | **N 400** | **N 400** | **N 400** | **N 402** |

*Data collection techniques for quality outcomes*

## MEASURING QUALITY

To measure quality, researchers at DM2 used their Qscore methodology, which leverages trackable, quality-oriented question sets used for many years to determine sample provider and respondent quality and characteristics. The longevity of these question sets provided data that gave significant benchmarks for the United States, from 50K+ interviews in the past year alone. In addition, some standard questions from sources such as the U.S. Census, were included to give a foundation for outside comparisons.

### *Provider rankings*

From a provider standpoint, we saw a range of scores that was wider than anticipated, running from very good (94) to suboptimal (86). This gave us a solid range of data to assess.

QUALITY SCORE * Provider

| Provider | Mean | N | Std. Deviation |
|---|---|---|---|
| A | 94.14 | 500 | 6.65 |
| B | 93.37 | 501 | 7.88 |
| C | 86.23 | 500 | 13.75 |
| D | 91.60 | 501 | 10.14 |
| **Total** | **91.34** | **2002** | **10.44** |

*Data collection techniques for quality outcomes*

### *Respondent rankings*

For respondents, Qscores were determined based on multiple standard engagement factors such as length-of-interview (LOI), time spent on grids, answer consistency, robustness of open-end responses, red herrings, and more. For this project, we found that overall quality for this sample was comparable to previous U.S. benchmarks using the same age/gender sample design. For example, quality was a bit lower for young males, and improved with age of respondent—holding true to typical industry patterns.

QUALITY SCORE * Quota Group

| Quota Group | Mean | N | Std. Deviation |
|---|---|---|---|
| Male 18–29 | 88.87 | 300 | 12.79 |
| Male 30–49 | 90.13 | 400 | 10.21 |
| Male 50+ | 95.00 | 300 | 5.77 |
| Female 18–29 | 89.15 | 302 | 12.74 |
| Female 30–49 | 91.46 | 400 | 10.69 |
| Female 50+ | 93.79 | 300 | 6.70 |
| **Total** | **91.34** | **2002** | **10.44** |

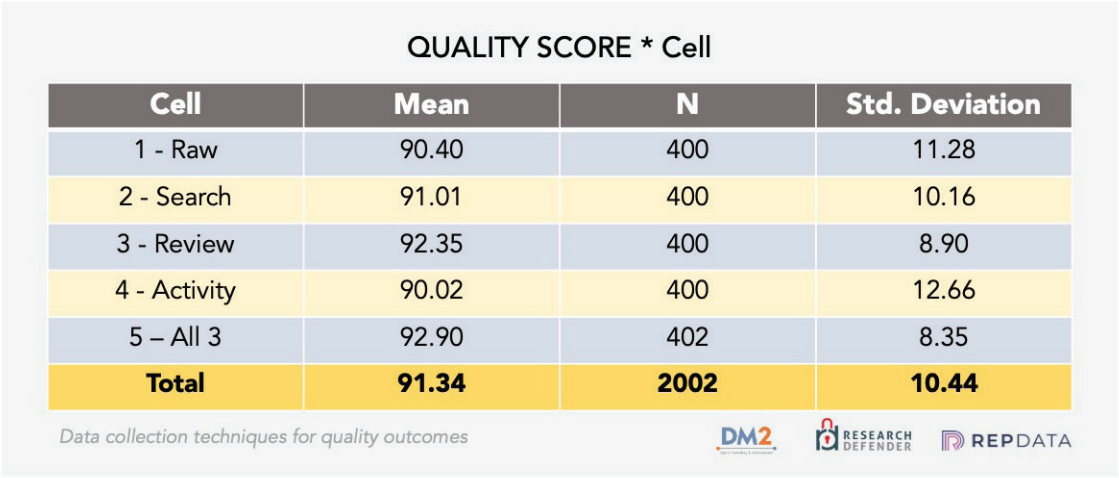
*Data collection techniques for quality outcomes*

## THE FINDINGS

### *Overview*

As indicated above, findings from the project were divided into five overall cells for data comparison.

1. Raw data: untreated data to provide a baseline for comparison.
2. "Search": data treated only with Research Defender's digital fingerprinting and fraud identification tool.
3. "Review": data treated only with Research Defender's respondent engagement and open-end response analysis tool.
4. "Activity": data treated only with Research Defender's respondent level tracking solution.**
5. Layered approach: data is subjected to methods two, three and four—all applied together.



QUALITY SCORE * Cell

| Cell | Mean | N | Std. Deviation |
|---|---|---|---|
| 1 - Raw | 90.40 | 400 | 11.28 |
| 2 - Search | 91.01 | 400 | 10.16 |
| 3 - Review | 92.35 | 400 | 8.90 |
| 4 - Activity | 90.02 | 400 | 12.66 |
| 5 – All 3 | 92.90 | 402 | 8.35 |
| **Total** | **91.34** | **2002** | **10.44** |

*Data collection techniques for quality outcomes*

As shown, using the layered approach improves data quality over untreated data cells at a statistically significant level (@ 99%). Using "Search" alone directionally improves quality, as does using "Review" on its own. While "Activity" doesn't show measurable improvements at this level (see footnote), it does help to remove the worst survey takers, as we see when we dig into the data.

***This segment appears as untreated data, to which we appended API information on respondent activity. The data was cleaned on the back-end, and exclusions were made during the analysis stage, rather than making the decision to remove the respondent in advance of the study.*

### *"Search": digital fingerprinting and fraud identification*

While using this method alone directionally improved quality (91.0 vs 90.2) in this study, it does provide insurance against catastrophic events, such as bot infiltration. This technique does provide insurance, in a way, by examining each respondent to see if they are part of a known network that is associated with a bot farm, or other fraudulent behaviors, to help eliminate them proactively.

While engaging new clients, Research Defender has learned of situations where up to one-third of a sample (or more) has been from bots. The example below shows how bot data can skew results in total, in a manner that might otherwise look reasonable but ultimately lead to inaccurate conclusions.

| | Stated Data | Percentage of Sample | Sample Size |
|---|---|---|---|
| Brand Awareness (norm) | 70% | 67% | 400 |
| Bots | 85% | 33% | 200 |
| >> Altered Awareness | 75%* | | |

*significantly different @ 95%

Data collection techniques for quality outcomes

### *"Review": Open-end response analysis*

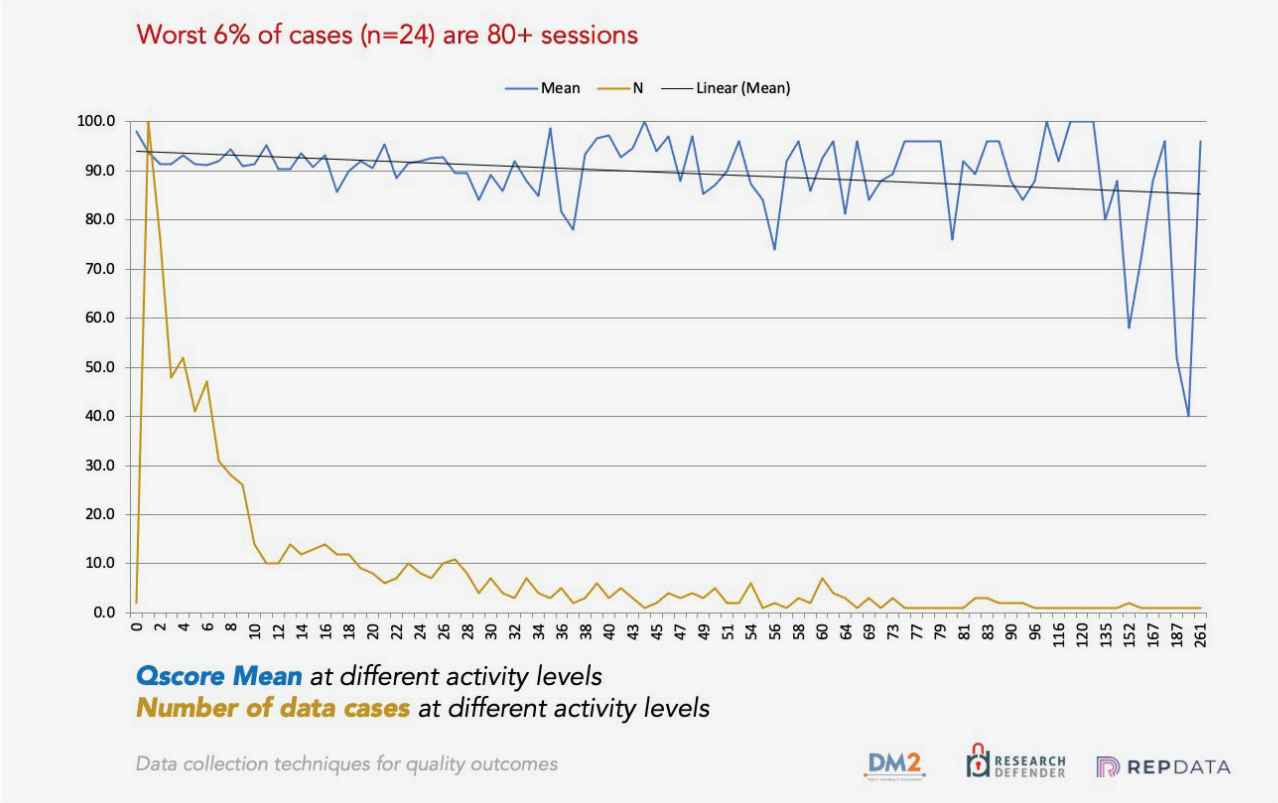
Our results showed using "Review" alone is the single-most effective approach to elevating quality, when using only one technique. This method examines any open-end responses checking for proper grammar, response length, profanity, copy/paste, and other attributes. It boosted Qscores significantly (@ 99%) in our study, from 90.2 to 92.4. This indicates that removing the most flagrantly poor open-end responders also improves closed-end data quality, decreasing the percentage of the most problematic respondents.

| | Cell | | | | | |
|---|---|---|---|---|---|---|
| | Total | Raw | Search | Review | Activity | All 3 |
| **Quality Score Tier** | | | | | | |
| Sample Size | 2,002 | 400 | 400 | 400 | 400 | 402 |
| **Excellent (100)** | | | | | | |
| Count | 640 | 127 | 112 | 132 | 131 | 138 |
| Column % | 32% | 32% | 28% | 33% | 33% | 34% |
| **Very Good (90-99)** | | | | | | |
| Count | 760 | 134 | 161 | 160 | 140 | 165 |
| Column % | 38% | 34% | 40% | 40% | 35% | 41% |
| **Good (80-89)** | | | | | | |
| Count | 377 | 86 | 85 | 71 | 62 | 73 |
| Column % | 19% | 22% | 21% | 18% | 16% | 18% |
| **Problematic (<80)** | | | | | | |
| Count | 225 | 53 | 42 | 37 | 67 | 26 |
| Column % | 11% | 13% | 11% | 9% | 17% | 6% |

Data collection techniques for quality outcomes

### *"Activity": Respondent level tracking*

We found that removal of the mostly highly active respondents (i.e., "professional respondents") by network does not necessarily ensure data quality improvements. When we eliminate the top 6% by activity level in our data set, Qscore improves only marginally from 90.02 to 90.48, significant at only 40% conf level (T-value of .516). While the most highly active do provide the worst data (with Qscores as low as 40), our findings suggest that they do, in fact, provide some good data. This indicates that some highly active IPs are well-organized in their approach to survey completion—reinforcing the notion of survey taking as a "cottage industry."



### *Layered approach: Review, Search and Activity together*

In the final cell for this study, we used "Search" followed by "Review" and concluded with "Activity" in our layered approach. This delivered a Qscore about 2.5 points higher than our raw data benchmark, providing a statistically significant difference at a 99% confidence level. Use of all three approaches also produced a sample with the fewest problematic respondents.

| | Cell | | | | | |
|---|---|---|---|---|---|---|
| | Total | Raw | Search | Review | Activity | All 3 |
| **Quality Score Tier** | | | | | | |
| Sample Size | 2,002 | 400 | 400 | 400 | 400 | 402 |
| **Excellent (100)** | | | | | | |
| Count | 640 | 127 | 112 | 132 | 131 | 138 |
| Column % | 32% | 32% | 28% | 33% | 33% | 34% |
| **Very Good (90-99)** | | | | | | |
| Count | 760 | 134 | 161 | 160 | 140 | 165 |
| Column % | 38% | 34% | 40% | 40% | 35% | 41% |
| **Good (80-89)** | | | | | | |
| Count | 377 | 86 | 85 | 71 | 62 | 73 |
| Column % | 19% | 22% | 21% | 18% | 16% | 18% |
| **Problematic (<80)** | | | | | | |
| Count | 225 | 53 | 42 | 37 | 67 | 26 |
| Column % | 11% | 13% | 11% | 9% | 17% | 6% |

*Data collection techniques for quality outcomes*

Our layered approach led to other measurable data quality improvements, including a consistently lower use of "None" and "Don't Know"—which are fairly typical survey responses when satisficing. In a series of seven importance questions, the use of "Don't Know" was minimal.

| | Raw | Search | Review | Activity | All 3 |
|---|---|---|---|---|---|
| **Avg** | **4.2%** | **5.6%** | **3.8%** | **6.4%** | **2.8%** |
| Q1 | 5% | 5% | 4% | 5% | 3% |
| Q2 | 7% | 13% | 8% | 9% | 4% |
| Q3 | 3% | 4% | 3% | 4% | 1% |
| Q4 | 3% | 4% | 3% | 6% | 1% |
| Q5 | 6% | 8% | 5% | 9% | 5% |
| Q6 | 3% | 4% | 2% | 7% | 2% |
| Q7 | 3% | 3% | 2% | 6% | 1% |

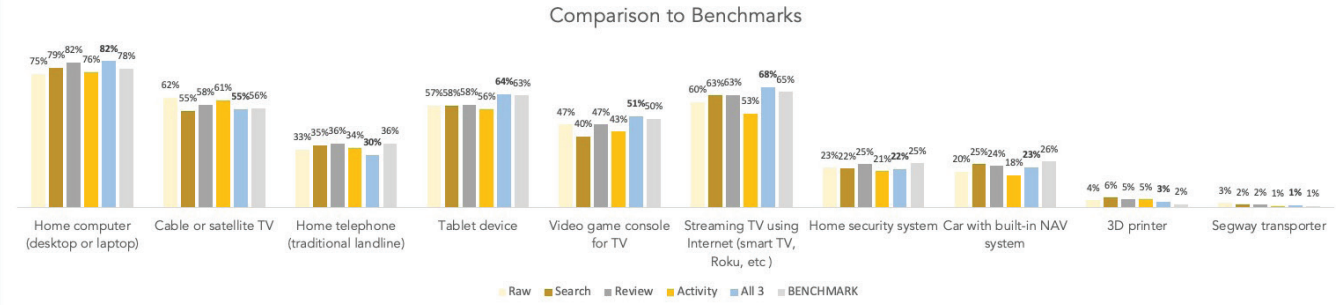
*Data collection techniques for quality outcomes*

Additionally, the statement of two fake items in lists (phone brand, social media app) was statistically lower than raw data (@ 95%) when using the layered approach.

| FAKE ITEMS | Raw | Search | Review | Activity | All 3 |
|---|---|---|---|---|---|
| **Aware Phone Brand** - Mentions | 10 | 11 | 8 | 8 | 1 |
| % Valid Cases | 3% | 3% | 2% | 2% | 0% |
| | | | | | |
| **Own Tech Product** - Mentions | 10 | 16 | 8 | 10 | 7 |
| % Valid Cases | 3% | 4% | 2% | 3% | 2% |
| | | | | | |
| **Use Social Media App** - Mentions | 6 | 7 | 4 | 7 | 2 |
| % Valid Cases | 2% | 2% | 1% | 2% | 0% |

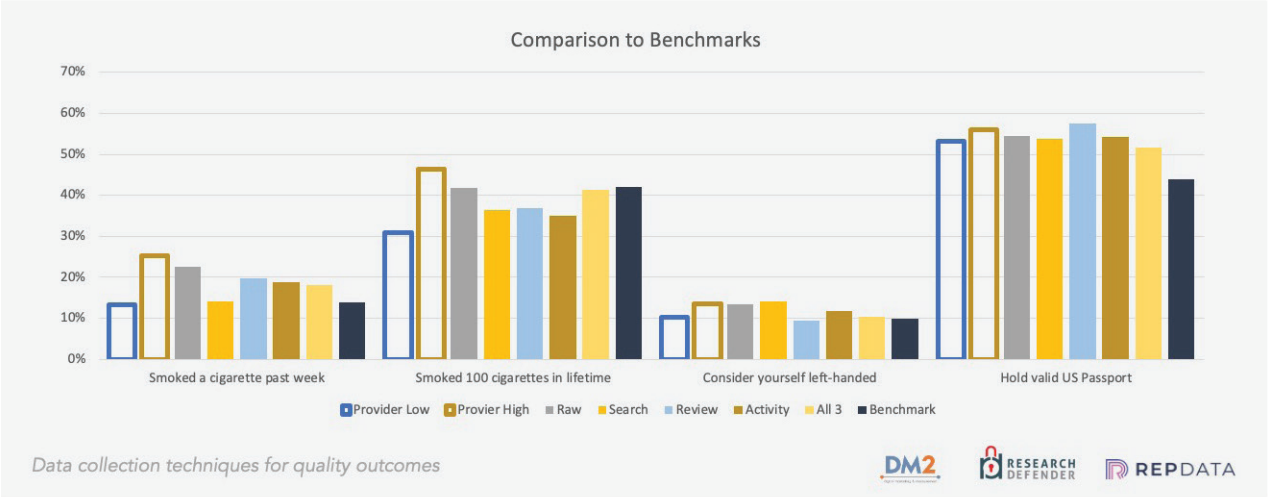
*Data collection techniques for quality outcomes*

When we examine the data against standard benchmarks (last updated in Q1 of 2021 from U.S. Census surveys, Statista, and other sources) we see that our layered approach of all three methodologies gets us the closest, with "Review" alone as the next best singular approach.




| | Raw | Search | Review | Activity | All 3 | BENCHMARK |
|---|---|---|---|---|---|---|
| Home computer (desktop or laptop) | 75% | 79% | 82% | 76% | 82% | 78% |
| Cable or satellite TV | 62% | 55% | 58% | 61% | 55% | 56% |
| Home telephone (traditional landline) | 33% | 35% | 36% | 34% | 30% | 36% |
| Tablet device | 57% | 58% | 58% | 56% | 64% | 63% |
| Video game console for TV | 47% | 40% | 47% | 43% | 51% | 50% |
| Streaming TV using Internet (smart TV, Roku, etc) | 60% | 63% | 63% | 53% | 68% | 65% |
| Home security system | 23% | 22% | 25% | 21% | 22% | 25% |
| Car with built-in NAV system | 20% | 25% | 24% | 18% | 23% | 26% |
| 3D printer | 4% | 6% | 5% | 5% | 3% | 2% |
| Segway transporter | 3% | 2% | 2% | 1% | 1% | 1% |

*Data collection techniques for quality outcomes*




## IMPLICATIONS

In our five cell study, the raw, untreated data (from Cells 1 and 4, Raw and "Activity") appeared most susceptible to lower quality, delivering the lowest Qscores and the highest percentage of problematic respondents. Cell 2 ("Search") has some impact on cleaning up the data, but appears best suited as insurance against "hijacked" datasets (via bots, organized fraudsters, etc.). As a singular approach, Cell 3 ("Review") is well-suited to improve data quality, raising scores more than other approaches alone.

***Our layered approach, using all three techniques, had these key benefits:***

- Increased Qscores significantly, while the % of Problematic respondents decreased
- Produced a dataset with the least satisficing behaviors
- Resulted in fewer "None" and "Don't Know" responses
- Delivered data closest to external benchmarks

Use of all three approaches mitigates risk of fraudulent data (with "Search" and "Activity") while producing a dataset of engaged respondents ("Review") who appear to answer both open and closed-end questions thoughtfully.

As with any data cleaning approach, users should be thoughtful in implementation, as heavy-handedness can inadvertently skew data through removal of portions of a sample that are required to adequately represent the true population.

## CONCLUSIONS

Delivering quality data requires an intelligent approach to data collection and fieldwork from start to finish. When seeking the best data quality from market research projects, here are some considerations:

- **Source broadly to minimize sample bias.** Find a sample partner that is not restricted to specific panels or sample suppliers which will give greater reach and diverse sourcing, ultimately providing a more representative sample for projects.

- **Find experts in execution.** There is no replacement for the "human" side of quality assurance. Better results can be achieved when projects are shepherded through the study process, including personal monitoring of sample and study quality and the creation of a post-project feedback loop to allow continuous improvement.

- **Use a blend of quality techniques.** Our research found that only by using a layered approach including techniques such as digital fingerprinting and fraud identification; text analytics for open-end responses; and respondent-level tracking across external behaviors were we able to achieve the best results.

## STUDY EXECUTED BY:

### Rep Data

Rep Data provides full-service data collection solutions for primary researchers, helping expedite data collection for primary quantitative research studies, with a hyper-focus on data quality and consistent execution. The company's mission is to be a reliable, repeatable data collection partner for market research agencies, management consultancies, Fortune 500 corporations, advertising agencies, brand strategy consultancies, universities, communications agencies, public relations firms and more.

**repdatallc.com**
**Email: reps@repdatallc.com**
**Phone: (817) 542-2520**

### Research Defender

Research Defender has created a secure platform to help clients take control of their traffic and quality of their product to create a clean, healthy and efficient ecosystem in the research industry. Having moved toward programmatic sample, a dynamic ecosystem that is constantly in flux, Research Defender exists solely to facilitate high quality and efficient transactions. As a truly independent entity, the company neither owns nor operates its own respondent panel or exchange.

**researchdefender.com**
**Email: info@ResearchDefender.com**
**Phone: (504) 298-9352**

### DM2: Digital Marketing & Measurement, LLC

DM2 provides clients significant capability in digital marketing, marketing research and business intelligence—capabilities centered around data to deliver quantifiable insight. Founder Chuck Miller developed the majority of DM2's products from experiences as a BI and Consumer Insights VP at AOL and Time Warner, focusing heavily on advertising metrics. Chuck and his team at his previous company, Digital Marketing Services (DMS), earned a reputation as pioneers and innovators in online marketing and marketing research. Thanks to this heritage, DM2 prides itself on innovation, continually exploring the ever-changing digital world to bring the best solutions to clients today.

**dm2corp.com**
**Email: info@dm2corp.com**
**Phone: (214) 505-5414**