

# The Devil Is in the Data

## A Respondent Classification Analysis Comparing the Best and Worst Survey Respondents

### Background

Starting in 2007 and continuing through 2008, DMS conducted a comprehensive research-on-research initiative, initially focusing on understanding the similarities and differences among multiple respondent surveying methodologies, including online panels, River (or real-time) samples, and RDD CATI interviewing.

The earliest phase of this research addressed key issues including respondent profiles, past and recent survey history, survey-taking motivations and earnings, in-survey behavior and data quality, and overall proximity to benchmarks based on the overall US population. The findings provided a clear understanding of the differences and similarities among respondent groups based on sampling methodologies and clarified how these differences might influence the research design and conclusions resulting from the use of a particular online sample.

Subsequent phases of this research focused on differences among specific sample types, with one phase concentrating on comparing eight different panel samples to each other, and another comparing five river or real-time samples. This research underscored the differences in quality among providers and uncovered how recruiting and sourcing biases can have an impact on data quality.

This final phase of the research combined the thousands of interviews collected over nine months to look at the data as a whole and segment respondents not based on their sampling methodology, but instead by their overall respondent quality. In this phase, we studied the different types of respondents that make up a typical data set and identified the characteristics of the best and worst respondents. A further consideration was to isolate the worst respondents and determine their impact on the quality of the data. Finally, we sought to understand how best to identify and remove those respondents who contribute to poor data quality.

Thus, the goal of this phase of the research was to call out the worst respondent – the devil in the data – whose presence in the data is known, but not always to what extent, and whose impact is not always quantified. This alone impacts the data quality because it is not always clear how pervasive it is, and researchers are often unclear how to identify the worst responders and uncertain what level of imperfection is permissible without affecting quality results.

### Research Design

Interviews were completed in three phases between December 2007 and August 2008. A total of 6700+ responses to the survey were received.

The research utilized the following sample sources:

- DMS River Sample (using DMS proprietary Opinion Place River Sample®)
- DMS Panel Sample (using DMS proprietary SurveySpree® panelists)
- “Real-time” samples from four competing sample providers
- Panel sample from eight competing sample providers
- CATI interviews (RDD)



Between 300 and 400 responses were collected for each sample source and statistical significance was tested at the 95% confidence level. With the exception of two river cells which were known to skew toward younger respondents, quotas were set to control for gender, age, income, and ethnicity, and were used to ensure that each sample source included in the original sample comparison resembled each other demographically and reflected the overall US population (according to US census estimates).

Precautions were taken at every step to ensure that the research was as unbiased, clear and objective as possible. In addition to surveying multiple different external river and panel samples, the sources of which were unknown to us, we also outsourced the data collection, coding, tabulation, and statistical analysis to further ensure objectivity of results.

The research questionnaire included the following key sections:

- Respondent profile: we surveyed respondents across demographic categories, attitudinal statements, product ownership, and a broad range of other topics.
- Respondent survey history: we measured the level and frequency of previous participation in market research, membership in research panels, and estimated survey earnings to understand how respondents from different sampling methodologies contribute to the overall body of online market research.
- Respondent survey behavior: we utilized standard data quality measures such as “data traps,” internal consistency, low incidence questions, survey completion times and open-ended responses to determine if the response quality of one respondent group outshines that of another. We also probed for motivations/detractors from joining panels, as well as motivations for participating in online surveys, to understand what moves respondents to accept a survey or panel invitation and if there is any underlying response bias.
- Benchmarks: we surveyed respondents from each sample source on a range of benchmark measures – from product ownership to presidential voting – to establish proximity to the overall population.

## Research Results

Unlike previous phases of this research, which focused on the differences and similarities among multiple respondent sampling methodologies, this phase of the research combined all of the respondents into one dataset to determine respondent quality segments and examine trends among the best and worst respondents.

Respondents were segmented into quality groups based on their responses to several quality measures that were either included directly in the survey as survey questions or captured indirectly in the survey meta-data. They included:

- **Speed.** The survey covered a wide range of topics, included more than 35 questions, and took the average respondent roughly 17 minutes to complete. Respondents who completed the survey in under 9 minutes were flagged for speeding through the survey.
- **Straight-lining.** The survey included two long grid-format questions; one was a 10-part grid, the other, a 16-part grid. Respondents found to have given the same value to each question (row) of the grid were flagged for straight-lining.
- **Internal Consistency.** The survey included three questions to test respondents’ internal consistency, that is, their ability to provide an answer to one question that did not contradict their answer to the same or similar question asked later in the survey. These included:
  - Kids in household - respondents were asked first in the interview screener and later in the closing demographics to report the number of children under 18 who lived in their household. Respondents who did not provide the same response to both questions were flagged for consistency.

- Item inconsistency (1) – respondents were asked two questions regarding their attitudes toward brand and price (*Price is more important to me than brand names / Brand names are more important to me than Price*). Respondents who agreed with both statements and respondents who disagreed with both statements were considered “inconsistent” responders. (Respondents who gave a Neutral value of “4” to either question were not considered inconsistent.)
- Item consistency (2) – respondents were also asked about their attitude toward their standard of living (I am perfectly happy with my standard of living / I’m not really happy with my standard of living). Again, respondents who gave both statements a positive or negative agreement rating were flagged for inconsistency.
- **Open-Ended Response.** This data flag captured the percent of respondents who provided a non-substantive response (such as “nothing” or “I don’t know,” etc.) to the open-ended question “What made you decide to take this survey today?”
- **Trick Question.** The first phase of this research employed a traditional data trap asking respondents to mark their place in the survey by selecting “3” on a 7-point scale grid-format question with multiple attributes. This trap was reworked in later phases to be clearer for respondents, where respondents were instead instructed to select “strongly disagree” on one attribute within a 13-part grid question. Respondents who did not comply with either instruction were flagged for failing the trap.

A simple segmentation of the overall data set based on these flags yielded the following four respondent quality groups:

- **Ideal Respondents** – these respondents have no data quality flags and make up an impressive 29% of the overall data sample. Demographically, Ideal Respondents tend to skew older and female, but are most alike in their response consistency and conscientious survey taking, including overall proximity to benchmarks, lack of patterned responses, thoughtful responses to open-ended questions, and careful attention throughout the survey.
- **Typical Respondents** – these respondents have one data quality flag and make up the single largest group of survey respondents with 41% of the overall data sample. Like Ideal Respondents, Typical Respondents are characterized by data that is in line with benchmarks and other evidence of careful survey taking (such as a low data trap failure rate, higher than average levels of internal consistency, and “sensible” or predictable responses to attitudinal questions). While they are not “perfect,” their one error is easily forgiven as their responses mirror those of Ideal Respondents: there is essentially low to no difference between Ideal and Typical respondents. For this reason, these respondents, along with Ideal respondents, make up the group hereafter called the **“Best Respondents”** and comprise the 70% of the sample whose responses can be trusted to be accurate and of the highest quality.
- **Imperfect Respondents** – these respondents have two to three data flags and make up a significant 27% of the overall sample. They tend to skew slightly younger and more male, and unlike the Best Respondents, who have in common a high level of consistency in their responses, the Imperfect Respondents are often inconsistent. At times, their responses are close to benchmarks and in line with average, while at other times, their responses appear to be outliers. Thus, while not always impacting overall data quality, these respondents can be inattentive and inaccurate, they are more prone to speeding through the survey, and tend to fatigue earlier than the Best Respondents.
- **Worst Respondents** – with four or more data quality flags, the Worst Respondents make up the smallest part of the data set. And these data flags – most often speeding, inconsistency, and straight-lining, contribute to data quality that is beyond suspect: it’s toxic. The Worst Respondents’ responses are often “off the charts,” with responses rarely comparable with benchmarks, higher than average rates of product ownership (even on low-incidence items such as hybrid cars), and contradicting responses to attitudinal questions.

The good news is that the Worst Respondents make up a very small part of the data – only 4% of the overall sample.

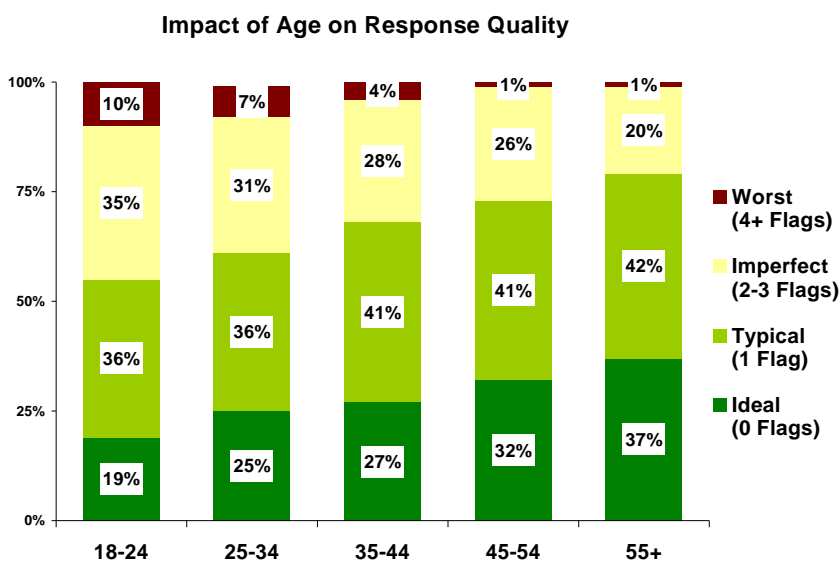
The bad news, simply put, is that you can't trust anything the Worst Respondents say. Speeding and straight-lining are incredibly pervasive among this group, at 73% and 40%, respectively, and likely earn the Worst Respondents enough other quality flags to place them in the Worst Respondent group. When reflecting on the Worst Respondents' impact on data quality, it is most disconcerting that they have failure rates as high as 80% on internal consistency questions, meaning that these respondents give contradictory answers to similar questions, or respond similarly to opposing questions. It is no surprise that the Worst Respondents are rarely in line with benchmarks, or that they give relatively improbable responses to simple questions (52% report that they own a laptop but 57% agree that "It would be fine with me if I never used a computer again").

The question then becomes not if or how often these devils sin, but whether they can be saved.

## Profile

Unfortunately, the Worst Respondents seem to be, at least on the surface, some of the research community's most sought after respondents: younger males are disproportionately represented in the Imperfect and Worst Responder groups.

However, we know from the data that the Worst Respondents do not always provide accurate or truthful information. While it is easily plausible that young males could race through a survey, committing multiple data quality infractions along the way, it is also possible for a respondent to pretend to be in a low-incidence group to qualify for a survey. Anecdotal data from this research and information from other research into data quality suggests that both scenarios (inattentiveness and fraud) are possible.



Age and gender aren't the only demographic skews among the worst respondent groups:

- Income may also be a factor, as respondents with household incomes less than \$25K are overrepresented in both the Worst and Imperfect Respondent groups, even when removing the youngest (and presumably least affluent) respondents. By contrast, the Best Respondents have a representative distribution of incomes.
- There are some differences by Race, as non-white respondents are overrepresented in the Worst and Imperfect Respondent groups.
- And finally, respondents with children in the household are overrepresented in the Worst Respondent group.

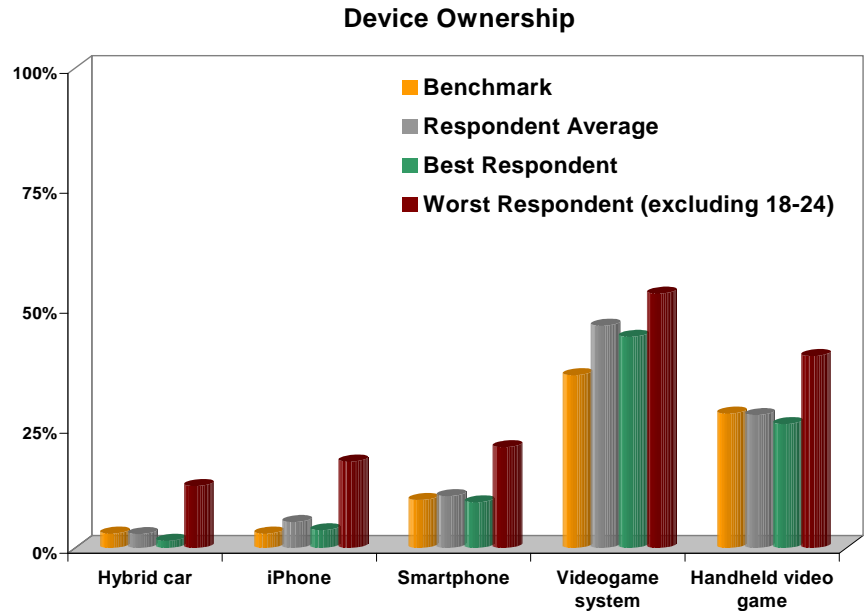
That several of the most coveted, hardest to attract demographic groups also turn out to be among the worst respondents suggests that it is extremely important to independently validate respondent details to ensure that, as far as information is verifiable, respondents are actually who they say they are.

## Impact

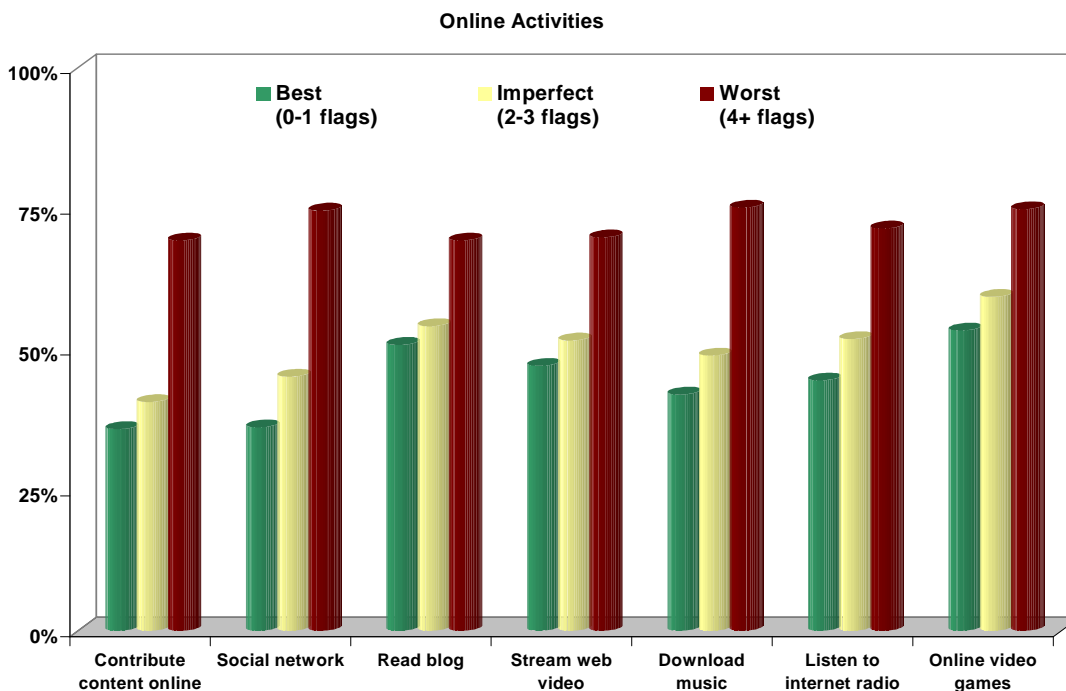
One of the biggest effects we see with the Worst Respondents is their tendency to over-state things – from product ownership, to online activities, to agreement with attitudinal statements. It's their contradictions and improbable responses (25% own hybrids?) that begin to plant the doubt that anything they say can be trusted.

One example is product ownership.

The Worst Respondents overstate product ownership on items ranging from phones to hybrid cars. Even when dropping the youngest respondents (who may arguably have higher rates of ownership of the latest technology gadgets like i-Phones and video games), the Worst Respondents still state device ownership rates that are significantly higher than the Best Respondent group and more importantly, the benchmark (based on industry data).



The Worst Respondents also raise doubts with higher than average participation in online activities. Due to their inaccuracy in other areas of the survey, it is difficult to say whether increased participation in online activities is a predictor of poor respondent quality, or whether the Worst Respondents are just over-reporting participation in online activities.



We could, however, theorize that if these Worst Respondents are in fact doing all of these sophisticated activities (such as social networking, blogging, downloading music, streaming video) to a greater extent than other Respondent groups, then the typical online interview must seem quite boring compared to the other content they consume. Much of today's Internet strives to engage all viewers but particularly the younger and sophisticated audience to visit more often and consume more content; and yet, when respondents visit research sites to complete an interview, they often see an interview environment not much different than that they would see on paper.

One possible impact on overall data quality is that product penetration is overstated and demand for services may not be as high as it appears to be. Because of their inconsistent answers to other questions, and in some cases their improbably high ownership of low-incidence products, it is likely that at least some of this data is inaccurate.

On the other hand, if respondents who are the most engaged with the most sophisticated content the web has to offer are also the internet survey's worst respondents then there is an opportunity (obligation) to create an environment equally as engaging for them to take surveys.

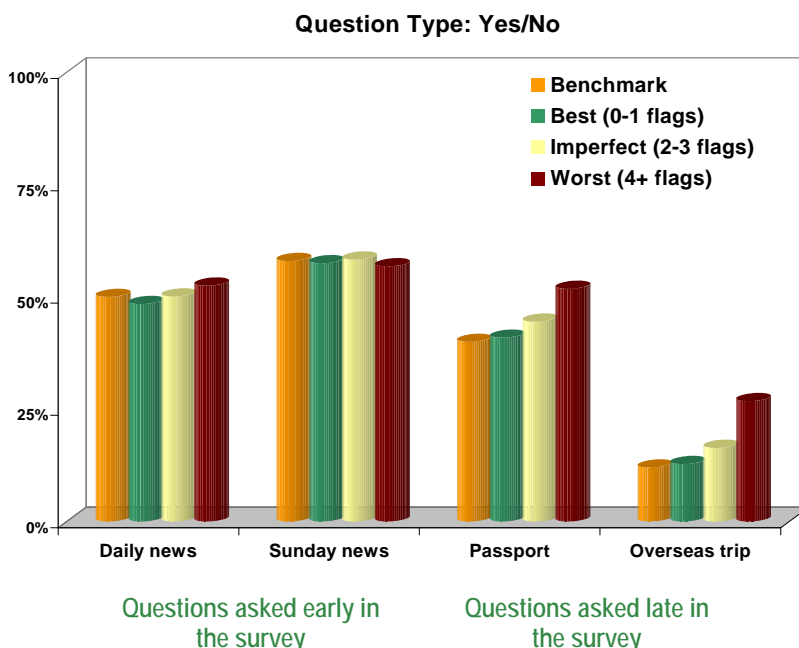
The Worst Respondents are also prone to satisficing, though in this survey that seems to be the result of straight-lining or their tendency to pick the first option available. Top box satisfaction with their cable / satellite television provider among those with four or more data flags is 53.9%, compared to 41% for the average and the Best Respondents. The Worst Respondents also under-report dissatisfaction (5% for respondents with four or more data flags vs. 14% for respondents with 0-1 data flags).

## Behavior

The survey data and respondent groups were analyzed by question type to determine whether the Imperfect or Worst Respondents performed better or worse depending on the relative complexity of the question (as well as its position in the survey), and more importantly, whether any of their data could be salvaged.

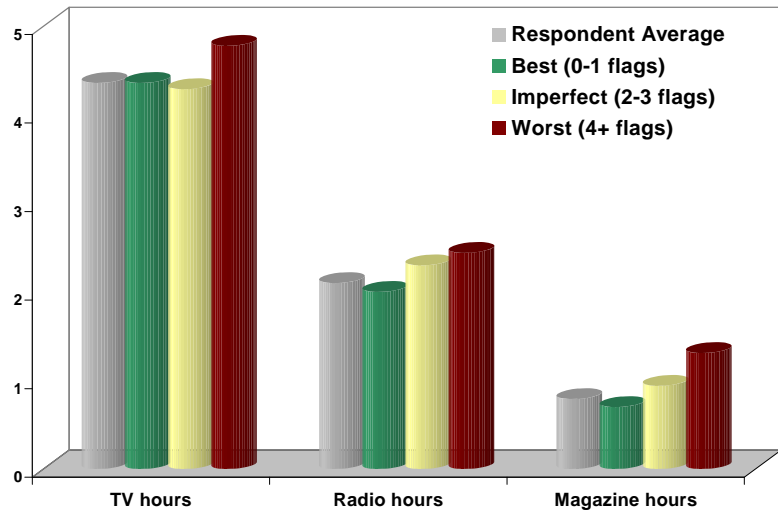
On the whole, the Worst Respondents' survey behavior seems to be almost equally bad, regardless of the relative ease or complexity of the survey question.

On several "Yes/No" format questions placed very early in the survey, the Worst and Imperfect respondents' responses are in line with average, raising hope that the worst respondents might be less troublesome on simple question formats. However, on several similar questions later in the survey, fatigue seems to be setting in, as both the Imperfect and Worst respondent groups start to stray from the benchmarks, to varying degrees: Imperfect Respondents are up to 37% higher than the benchmark, while the Worst Respondents are almost double the benchmark on the question of whether they have traveled overseas in the last year.



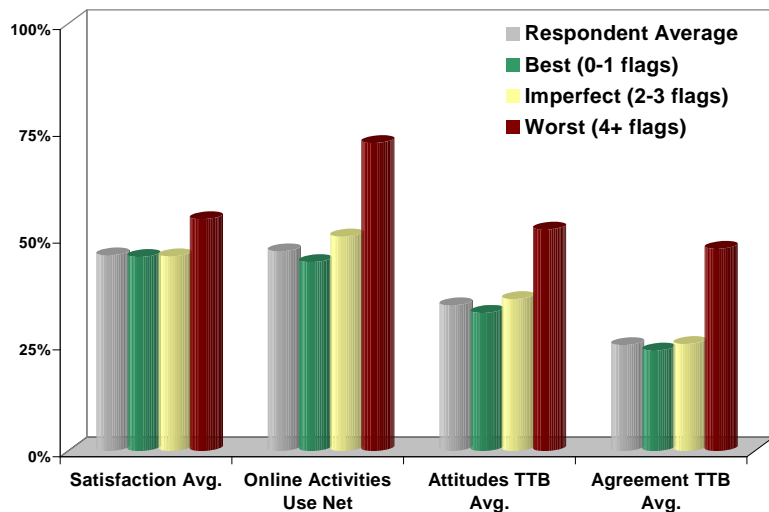
On open numeric response questions asked early in the survey (shown in the chart to the right), the Worst respondents are outliers, posting numbers 10% to 66% higher than the respondent average. Imperfect respondents are at least within acceptable proximity to the majority. On open numeric response questions asked later in the survey (including estimated earnings from market research), several of the Worst respondents posted outlandish numbers (\$10,000+), making their average earnings numbers exponentially higher than the best and average respondent.

**Question Type: Open Numeric Response**



While grid questions can be difficult for even the most conscientious respondents, they are almost wasted on the Worst respondents, given the quality of data that comes out of them. Grid-format questions are where straight-lining and patterned responses come into play, and where the Worst Respondents are so far outside of the normal range of responses that it is difficult to believe that their responses are an accurate reflection of their satisfaction, frequency of use, beliefs, or attitudes. Though it is possible to set a simple data trap or warning to catch the straight-

**Question Type: Grid**

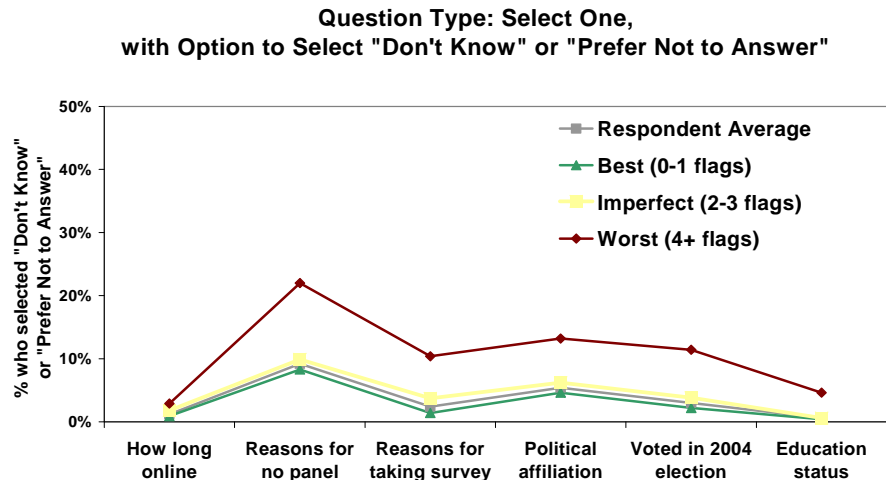


liners, based on the behavior of the Worst Respondents, it's difficult to believe that even a simple data trap would cause the Worst respondents to respond to these questions more carefully. However, while the Worst Respondents seem more and more like outliers as grid length and survey time increase, the Imperfect respondents are still within an acceptable range, with a greater use of the scale than the Worst respondents, better scores on the grid-format consistency questions, and "sanity check" passes (such as low agreement that they could live without a computer).

*When they're not giving bad data, they're giving no data.*

It's clear that the Worst Respondents take shortcuts throughout the survey, from speeding to straight-lining to providing careless responses on open-ended questions. Thus it is not surprising that on any questions where they had the opportunity to respond "don't know" or "prefer not to answer," the Worst Respondents were quicker than others to take that opportunity.

In fact, the Worst respondents refused the opportunity to answer some questions six times as often as the Best Respondents. Though overall refusal rates are relatively low even among the Worst Respondents, this behavior, combined with their repeated tendency to provide inaccurate or improbable data, reinforces that their entire data set is suspect and should be discarded.



## Remove and Discard

Because it is their *behavior* rather than any specific set of characteristics that brands them as the Worst Respondents, it is difficult to implement what would be the ideal solution to address bad respondents -- successfully screening them out of the survey entirely before the survey even starts. Thus, one option for dealing with the Worst Respondents and improving overall data quality is to identify the Worst Respondents post-survey based on specific behaviors and then remove their responses entirely from the data set.

The obvious goal of the data cleansing process is to strike the right balance of eliminating bad respondents while keeping good respondents. Accomplishing this is a function of selecting the right questions on which to base the decision while allowing some level of forgiveness for respondents who make honest mistakes. The number of quality flags to apply is critical: too many, and we risk muddying up the survey with multiple quality traps; too few, and we risk not catching and removing enough potentially toxic respondents. Selecting the right questions is also a consideration: researchers have been using traps and data quality measures for some time to improve the quality of the data set, but it's important to exercise caution in how and which of these we apply, as some traps and tricks can trip up even acceptable respondents.

An analysis of the sample by the eight data quality flags used in the survey determined that there are several flags that, when used in combination with each other, can effectively remove the worst respondents without losing too many acceptable respondents.

A consistency pair within a grid question is one good way to test respondents' attentiveness, but the key is for the concept to be simple enough to follow that a careful survey taker would not be confused by it.

Consider, for example the following consistency pair:  
*Price is more important to me than brand names*  
*Brand names are more important to me than price*

While the "brand/price" consistency question captures a very high number of the worst respondents (84%), it also captures 41% of the acceptable respondents. It is clear that even careful survey takers can be confused by long and involved attitudinal or belief statements like this, especially at the end of the survey.

On the other hand – a simpler consistency pair involving agreement with attitudinal statements about their standard of living captured an impressive 80% of the worst respondents while only tripping up 17% of the acceptable respondents.

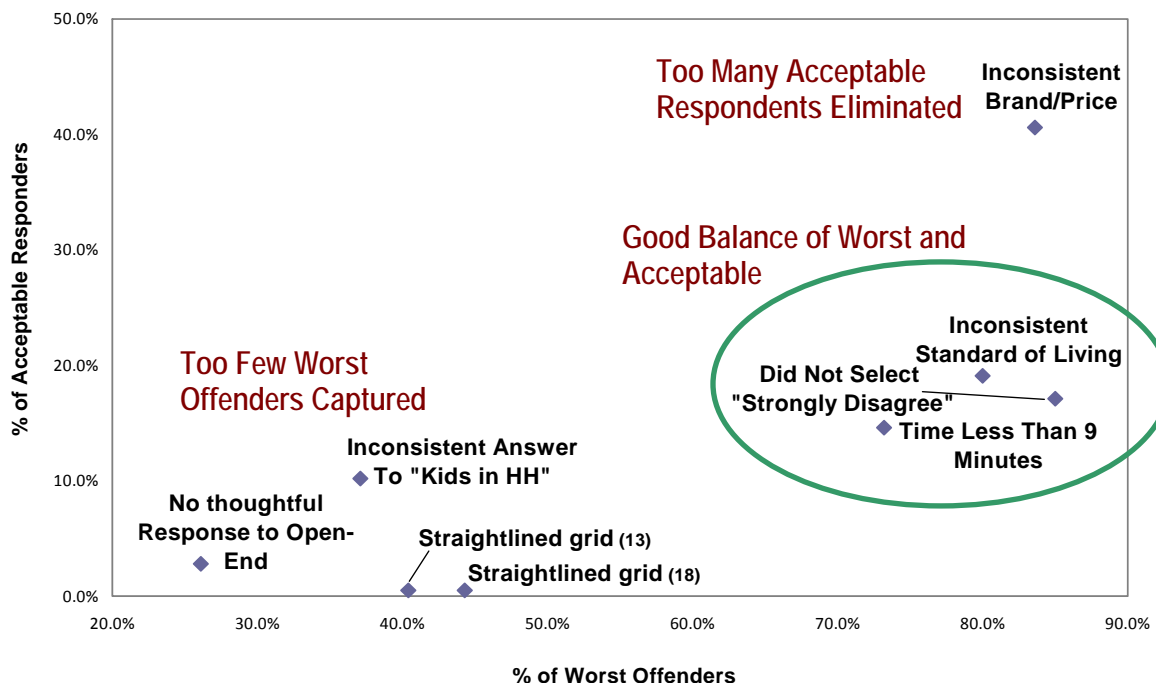
*I am perfectly happy with my standard of living*

*I am not really happy with my standard of living*

	% of Total Sample	% of Worst Respondents	% of Acceptable Respondents
Inconsistent on Kids in HH	11%	37%	10%
Time less than 9 minutes	17%	73%	15%
No thoughtful response on open end	4%	26%	3%
Did not select "Strongly Disagree"	20%	85%	17%
Inconsistent on Standard of Living	22%	80%	19%
Inconsistent on Brand/Price	42%	84%	41%
Straight-lined 13-part grid	2%	40%	1%
Straight-lined 18-part grid	2%	44%	1%

The "trick" question (*to ensure data validity, please select "strongly disagree" for this row*) is good, but cannot be used in isolation, as it is especially confusing for first time or casual survey takers who are not as familiar with survey instructions.

Getting rid of the most egregious speeders is also a simple but effective way to identify and eliminate the worst responders, but it must be used in combination with the other quality tools to ensure that it doesn't unfairly snag those who are actually just familiar enough with research to be able to complete surveys more quickly than average.



If the best balance is to remove the worst respondents, screening for and eliminating respondents with two of those three flags would remove 92% of the worst respondents and only 6% of the acceptable respondents for an estimated total of 10% of the sample.

If 10% seems too much, screening for and eliminating respondents with all of the three above would successfully remove half of the “Worst Respondents” and only .2% of the acceptable respondents, for a total of only 2.2% of the sample.

## Conclusions

After evaluating over 6,700 interviews from an incredibly long and complex survey and then using a combination of eight data flags to determine respondent quality segments, it is inspiring to see that most respondents are inherently good. The largest part of the sample (70%) had one data quality flag or less over the course of 40 questions. Another group of respondents had two to three data flags, but their proximity to benchmarks and consistent responses suggest that they are for the most part engaged with the survey process and trying to provide genuine and truthful responses.

Unfortunately, a small percent of the sample (4%) is inherently bad. The Worst Respondents’ behaviors seem to compound each other (speeding and straight-lining lead to inconsistency and failing traps), though their issues seem to be beyond what simple error checking could catch. They fail to provide quality data on almost every level – they deviate from benchmarks, they rate satisfaction and product use more than 100% higher than the average, they refuse to answer questions at up to six times the rate of the average respondent, and their high levels of failure on trick questions and consistency pairs suggest that they do not read or answer survey questions carefully enough to accept that ANY of their responses are accurate or true. In short, their data is toxic to the quality of the overall data set and as much as possible should be removed.

Going forward, we suggest the following quality guidelines :

- Make the survey environment more engaging.

While the Worst Respondents are beyond repair, there are other respondents in our surveys that do provide accurate information, even if they make more than an acceptable amount of mistakes. These respondents can fatigue earlier in the survey than the most attentive respondents, and can get tripped up on long grids with complex or confusing statements or beliefs that they must evaluate.

- Use a more selective data cleansing process (post survey) that relies on a combination of triggers.

We know that using one flag injudiciously can just as easily punish an acceptable respondent who makes an innocent mistake as the worst offender who makes many mistakes, but also that most surveys don’t have the time or ability to include eight data traps to capture only the worst offenders. But using two or three select data quality flags can greatly improve the odds of removing only the worst respondents without punishing everyone else, and as one of these is not even a question, it’s a manageable number to insist on including in each survey.

- Speed
- Well-worded and easy to understand “trick question”
- Well-worded and easy to understand consistency pair
- Don’t know / refusals

- Be prepared to discard 2-5% of your sample.

Researchers must be prepared to remove a small percentage of the data set in order to improve the overall data quality.

## Suggestions for Further Research

As a result of this research initiative, we know more about the worst responders that we ever have. However, respondents are evolutionary and data quality is a continuum, meaning that as research changes and participants learn, what we should look for will change. DMS will continue to analyze respondents using similar and new techniques. Additionally, we plan to do more work on combining Optimus (computer identification) and Idology (personal identity validation) to weed out the liars up front and remove them, and then see what types of people are typically removed. Once undesirable participants are identified and declined, we will continue to examine what it means for hard-to-reach respondent segments in terms of feasibility and how people respond to being “caught.”

Additionally, we will conduct research among hard-to-reach respondent segments, such as young males or ethnic groups, that cannot be validated, and compare them to those that can be validated. The plan is to determine if there are any differences in quality among the two groups to prove whether certain demographic groups are inherently worse respondents or actually liars attempting to get into the survey by saying they are age, gender and ethnicity of quota groups that are more likely to qualify.

For further information on this and other research on River respondents, please contact:

**Melanie Courtright** | Vice President | 214-222-6176 | [m.courtright@corp.aol.com](mailto:m.courtright@corp.aol.com)

**Denise Brien** | Senior Research Manager | 703-265-1237 | [denise.brien@corp.aol.com](mailto:denise.brien@corp.aol.com)

**Marjette Stark** | Senior Vice President | 703-265-0262 | [marjette.stark@corp.aol.com](mailto:marjette.stark@corp.aol.com)