••• data quality

Fight the good fight

How to find and eliminate cheaters, liars and trolls from your surveys

| By Joe Hopper

snapshot

The author offers 10 things you can do to safeguard your data from the effects of undesirable respondents.

A recent RFP we received specified this: "Your proposal must include a detailed description of control measures in place to understand the validity of respondents." That is a smart client. They know that many surveys are now plagued with fraud in many of the same ways that social media is overrun with bots and trolls.

We welcomed the opportunity to address the issue for two reasons. First, because it is a serious problem that every organization collecting survey data needs to understand and guard against. If you or your vendors have not taken measures to find and eliminate bad data in every survey you conduct, your leadership is making bad decisions based on your bad data.

The second reason we welcomed the opportunity is that we have worked hard over the years to develop good protocols to find and eliminate cheaters, liars and trolls. And we are proud of how effectively we do it. Every time we work with sample suppliers, partners or clients to remove and replace bad data, we are reminded that not many companies bother, which honestly boggles my mind.

Our protocols are not secrets and they're not rocket science. They are just part of doing diligent, rigorous and careful research. Here are the most important things we are doing. We think you should be doing them, too.

1. Build an elaborate screening path. Bad survey respondents know that most surveys target specific buyers, or age groups, or decision makers with unique qualifications. And they know how to game their responses (and lie) to get in. They succeed because survey designers make it easy and obvious. So do this instead: Build a series of several, somewhat complex, screening questions. Allow for multiple responses that will conflict with each other if someone is answering randomly or if they are selecting many options to get in. You will see your qualifying incidence drop dramatically. That's a good thing.

2. Avoid river sample. If you can, that is … and for now, until problems of quality control and identity verification are solved. Most sampling panels use double opt-in verification to confirm that the people they are inviting into surveys (and compensating for their time

www.quirks.com/articles/2019/20190108.aspx and effort) are real, individual people. But they also unfortunately augment with real-time recruiting through ads and online pop-ups. There is no telling who (or what) gets routed into your survey and there is no tracing back to validate that they were real. If you're running high volume, very fast and very cheap surveys, chances are you are getting river sample. A lot of it is probably bad data.

3. Make rule-based cuts. A cardinal sin of research is cherry-picking data. Cleaning out poor-quality or fraudulent data can veer dangerously close to cherry-picking if done ad hoc. Do not scan through data manually looking for weird respondents. Rather, come up with rules you apply programmatically to all data. For example, decide ahead of time what counts as an unusually large or small numeric entry or what counts as straightlining or speeding. As you are deciding on who to cut and how many, never look at how your decisions will affect the outcome of your survey results, as this becomes the very definition of cherry-picking data.

4. Build tiers of red flags. To apply rules programmatically, write syntax that flags every instance of suspicious respondent behavior. "A" flags mark data that will result in automatic removal (like coming from a known fraudulent address). "B" flags are for serious violations, like implausible answers that contradict other data. "C" flags are for softer violations like speeding or inattentive behavior. Decide how to apply cuts based how many flags you see and in what combinations you see them. One or two C flags are okay and you can probably keep those respondents. But multiple flags, especially if they are B flags, signal bad data for cutting.

5. Include an open-ended question. Make sure it is a (required) question that everyone gets and that everyone will be able to answer it thoughtfully. At the end of your survey, review every response to evaluate whether it has thoughtful content. Bad respondents give you bad answers. Some will key-smash with random letters. Some will cut and paste sentences or paragraphs from other sources, even from your own survey. Some type in irrelevant information or completely generic-sounding answers that don't answer your question. Tag these responses with A, B or C flags based on how seriously bad they are.

6. Review IP addresses. When you start flagging and cutting specific respondents for quality problems, take a look at their IP addresses. You will probably see many of them coming from similar addresses. If you use an IP-lookup tool, you will also notice that many are from rural or foreign locations with weird names like Huge Data Network LLC. They look fishy and they are. Cut all respondents with those IP addresses. Then permanently block those IP addresses from your current survey and all future ones. Sample providers will say that they are doing this for you but trust me, they are not.

7. Build in quality checks. Quality-check questions have fallen out of favor because panel providers are convinced that "inattentiveness" is normal and often the result of poor survey design. They are partially right. But if you're like us, you almost never design long and tedious surveys that would explain inattentive behavior (most companies unfortunately do). We find that the overwhelming majority of respondents who fail quality-control questions fail our other quality-control checks as well. So go ahead and include them. They are a useful means of triangulating bad data so you have a solid rationale for who to cut and who to keep.

8. Look for inconsistencies. For some survey questions you may be tempted to restrict the logic of possible answers to make back-end data cleaning easier. For example, if you ask how many years ago a person was diagnosed with a disease, why not forbid entering a number that is greater than their age? Because questions like these give you ideal opportunities to validate the credibility of respondents, that's why. There are usually several questions in a survey that will elicit logically consistent responses if respondents are telling the truth. Lay out all the possible contradictions you can find, then check and flag each one for every respondent who provides inconsistent answers.

9. Review time stamps. Decent survey platforms will record the "time in" and "time out" of every person who takes, or attempts to take, your survey. You should download and keep that data along with all the important stuff. Calculate how much time each respondent spends in your survey. Very long times are infrequent and usually OK; it means somebody got interrupted and resumed taking the survey later on. Very short times are not OK; it means somebody raced through, clicking answers without reading the questions. Multiple, sequential time stamps can also reveal clusters of survey attempts (and often successful completes) from robots or fraudsters that should be flagged for removal.

10. Search for patterns. We try to avoid too many grid-style questions in our surveys (opting for stand-alone questions instead) but grids are often better and they can be an excellent way to find people who are not taking surveys seriously. Straightlining is when a respondent clicks the same answer for all questions in a grid. Sometimes it's legitimate and sometimes not, so decide ahead of time which grids to analyze for straightlining. Search for unlikely patterns in other questions as well, like sequential numbers in numeric entry boxes. Unlikely patterns should be flagged as indicators for potential cuts.

Provide thoughtful input

I feel somewhat bad writing this article, worried that some might conclude we should be wary of the people taking our surveys. But that's not true. The vast majority of survey respondents participate in good faith and we can see in their responses genuine efforts to provide thoughtful input to our questions. Yes, we see it every day, so thank you dear respondents! It's that very small slice of bad actors (who cheat and cheat again and magnify their efforts through technology as well) that we're after.

Opinion polls and surveys work amazingly well (and can help you make better decisions) because good people want to share honest opinions – and they do. The key is to ensure your analysis and conclusions are based on their honest opinions by outsmarting the cheaters, liars and trolls who may be messing you up.

Joe Hopper is president of Chicago-based Versta Research. He can be reached at jhopper@verstaresearch.com.