www.norc.org | info@norc.org

Research Brief Series #5: The Use of Data Cleaning Procedures in Probability-Based Panels

Prepared by Zoe Grotophorst, Ipek Bilgen, and David Dutwin

Users of online panel data are often concerned about potential data quality issues introduced by fraudulent, "professional," and/or inattentive respondents. There are no widely accepted industry standards for cleaning online panel data to remove suboptimal responses from the data or excluding respondents who are chronically providing suboptimal responses from the panel. In this white paper, we describe the issues concerning suboptimal response data and our recommendations and procedures to ensure high quality data.

Generally, data cleaning is concerned with a number of possible deleterious behaviors by survey participants. These include "straightlining," which happens when a respondent provides the same answer to each question in a grid; "skipping," which involves a respondent failing to answer a sufficient number of survey questions; and "speeding," where the time it takes a respondent to complete a survey would not be feasible; and "lagging" or "laggards," which involves respondents that spend too much time taking a survey (especially on surveys designed for single-session administration). All these behaviors are indicators of suboptimal responses (AAPOR, 2010). All surveys, probability or nonprobability, panel or cross-sectional, must be concerned with respondent behaviors that can compromise data quality.

The Special Case of Nonprobability Panels

Before discussing the impact of straightlining, skipping, speeding, and lagging on probability panels and efforts to mitigate their effects, we first note that there are additional data quality concerns in panels, but which are limited to nonprobability panels. In nonprobability-based panels specifically, there is a substantial risk of self-selection bias because such panels are built without a statistically valid sample frame. Instead, builders of non-probability samples use a variety of opt-in methods to advertise the panel (e.g., websites, pop-up ads, or social media) to potential panelists. Successful recruitment often relies on social media and other web users seeing the web ads and then self-selecting into panel membership. In contrast, probability-based panels have precise control over who is invited and allowed to join a probability-based panel. Probability panels use traditional sample frames, most commonly the address-based file maintained by the U.S. Postal Service, for selecting sample units with known probabilities.

Self-selection process for nonprobability-based panels creates an opportunity for another issue, survey fraud. The literature defines survey fraud in a variety of ways, including respondents who complete multiple surveys manually or through automation (e.g., bots), misrepresent themselves when they sign up for panels, and/or misrepresent themselves in screening questions qualifying them for individual surveys (Teitcher et al, 2015; Hulland and Miller, 2018; Le Guin, 2005; AAPOR, 2013). To protect against fraudulent activities, AAPOR (2010) recommends validating each respondent's identifying information at enrollment (e.g., name, address, IP address, phone, email) against third party sources, examining responses for respondents who choose all options in multiple response qualifiers or give low-probability

answers, and performing consistency checks. While these measures are critical to ensure data quality in nonprobability-based panels, they are not applicable to probability-based panels because of such panels' reliance on statistically valid sample frame and rigorous selection procedures for sampling households. Recruitment efforts for probability-based panels are focused on locating and recruiting only the specific individuals selected from a set sampling frame. Researchers use various traditional and well-established methods to invite them into the panel, including contacting the sample members by mail, email, phone, or in person.

The ability to self-select into nonprobability-based panels increases the number of "professional respondents," people who belong to many panels and complete large numbers of surveys. Various studies have shown that nonprobability-based panelists typically belong to multiple panels (Tourangeau, Conrad, & Couper, 2013; Craig et al., 2013). While recent research shows professional respondents may not provide sub-optimal data (Zhang et al., 2019; Hillygus et al., 2014), their membership in multiple opt-in panels can create a data quality issues. If a survey uses multiple panel sources blended into a single sample, some people may be represented more than once and could submit multiple surveys (Walker, Pettit, and Rubinson, 2009). To mitigate this issue, non-probability panels must carefully de-duplicate panelists at enrollment and researchers using blended samples must deduplicate prior to sending out a survey. Probability-based panels do not need these data cleaning measures because their panel members are randomly selected from the targeted population (e.g., general U.S. population) and then invited to join. Professional respondents are no more likely to be present in a probability-based sample than any other person in the population. Unlike a nonprobability-based panel, they cannot opt-in.

Data Quality Issues Common Across All Types of Panels

As noted earlier, universal concerns for suboptimal data comes from four measurable behaviors: "straightlining," "speeding,", "skipping," and "laggard" behavior. The AAPOR Task Force (2010) notes that procedures for identifying inattentive respondents as the "most controversial" of all data cleaning procedures, but researchers generally agree on certain best practices: looking at response patterns for those who straightline or randomly select answers in grid questions; reviewing time stamps to identify respondents with especially short survey completion times or abnormal dwell times; reviewing the number of questions with non-substantive answers or refusals; and reviewing verbatims or open-ended numerical text box responses for plausibility, internal consistency, and copy-paste errors (Greszki et al, 2014; Malhotra, 2008; AAPOR, 2010; Le Guin 2005; Bertoni, 2022). That said, it is not always clear where the line should be drawn whereby survey responses from a respondent should be considered suboptimal or even fraudulent. There are many situations to a straight line of answers to a grid of questions is perfectly valid. As well, speeding is related to respondent demographics (Yan and Tourangeau, 2008) and item design (Heerwegh, 2003; Tourangeau, Couper, & Conrad, 2004). "Don't know" responses (the principal way in which a respondent "skips" providing a substantive response) may originate from the desire to make the most honest or accurate choice (AAPOR, 2010). Finally, researchers are advised to be aware that excluding respondents based on certain behaviors may worsen systemic bias. Straightlining, for example, has been shown to be more common among lower education respondents (Kim et al, 2019). It is therefore important that decisions about exclusion be made carefully, taking multiple available data quality measures into account, as well as the particulars of the survey and target population.

Suboptimal Responses and Respondents

To understand suboptimal responses and respondents, we analyzed NORC's AmeriSpeak® conducted a deep analysis of suboptimal responses in 2022 and found that firstly, suboptimal behavior is highly isolated. Just over 10 percent of panelists have ever speeded, as defined by a survey length two times fast than the median length of a given survey (a conservative estimate). But of these, only 4 percent have speeded more than twice despite participating in many AmeriSpeak surveys over time. Even fewer have skipped more than half a survey completed: less than two percent of panelists have ever done so, and less than one percent have done so more than once. Straightlining is a more difficult assessment, again given that it is always technically possible to provide fully valid answers in a straightlining pattern, but we see similarly

low frequencies to this behavior through more comprehensive reviews during data processing.

The Center's Perspective

In contrast to cross-sectional surveys, panels have the advantage of having some level of a relationship through ongoing communication with their panelists. Again as one example, AmeriSpeak strives to make this relationship personal, providing emails that update panelists on panel developments and results from recent studies, a vibrant helpdesk of live interviewers to help assist panelists at any time, and even by sending well-wishes for birthdays and other occasions. It is important to value panelists and aim to provide a beneficial and pleasant experience to them so that they also value being an panelist. Further extending this point, we understand that there are known question formats that are at higher risk of discouraging attentiveness. The risk of straightening, for example, can be in part managed by following best practices in survey questionnaire design, even in nonprobability samples, by avoiding long grid questions, putting in "speed bumps" that slow down the pace of survey administration, effective transition screens for maintaining the morale of the respondent, and so on.

Beyond the general practices noted above, the AmeriSpeak response to suboptimal behavior has been twofold. First, if a panelist is flagged as having suboptimal responses due to straightlining, skipping, or speeding, such data is cleaned from the survey. In anticipation of some such cases on any given survey, we ordinarily gather a few extra cases in every survey so that we do not fall under the requirement minimum number of surveys for any given project. Second, we reach out to respondents who have likely provided suboptimal responses. We are currently experimenting with different interventions to assess the most effective means to improve survey responses in the future. To date, we have found no panelists whose behavior is so chronically suboptimal as to suspend them from panel surveys. We have however at some point contact just under four percent of panelists to encourage them toward more consistently high quality behavior, by describing the importance of good data, how as a panelist they can help, and how much we value them as panelists. NORC as well has implemented other strategies experimentally to assess the most optimal methods by which to improve data and panelists' experience with the panel.

Probability panels do not experience problems particular to nonprobability panels such as fraud due to bots, of the existence of professional respondents that belong to many panels. Panels like AmeriSpeak have rigorous standard protocols that flags cases that speed, straightline, and/or skip. Most will discard surveys that exceed standards and will regularly communicate with panelist to be transparent and encourage them to provide the most thoughtful responses possible. Overall, we believe as a result of such efforts, instances of data failures due to poor respondent behaviors will be low, but we believe it is critical to always review suspect cases and provide replacement interviews whenever necessary to guarantee high quality data.

References

American Association of Public Opinion Researchers (AAPOR). Task Force Report on Non-Probability Sampling. 2013.

American Association of Public Opinion Researchers (AAPOR). Task Force Report on Online Panels. 2010.

Bertoni, N. "Evaluating Data Quality in Online Panels with a Focus on Individual Respondents." (Current Innovations in Probability-based Household Internet Panel Research Conference, Virtual, March 4, 2022).

Craig, B.M., R.D. Hays, A.S. Pickard, D. Cella, D.A. Revicki, and B.B. Reeve. "Comparison of US Panel Vendors for Online Surveys." *Journal of Medical Internet Research* 15: no. 11 (2013)

Downes-Le Guin, T. 2005. Satisficing Behavior in Online Panels. MRA Annual Conference and Symposium.

Greszki, R., M. Meyer, and H. Schoen. The Impact of Speeding on Data Quality in Nonprobability and Freshly Recruited Probability-Based Online Panels. April 11, 2014. Wiley Blackwell.

Heerwegh, D. "Explaining Response Latencies and Changing Answers Using Client-Side Paradata from a Web Survey." *Social Science Computer Review 21*, no. 3 (January 1, 2003): 360–73

Hillygus, D., Jackson, N., & Young, M. *Professional Respondents in Nonprobability Online Panels*. April 11, 2014. Wiley Blackwell.

Hulland, J., and J. Miller. "Keep on Turkin'?" Journal of the Academy of Marketing Science: Official Publication of the Academy of Marketing Science 46, no. 5 (September 1, 2018): 789–94.

Kim, Y., J. Dykema, J. Stevenson, P. Black, and D.P. Moberg. "Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys." *Social Science Computer Review* 37, no. 2 (February 20, 2018): 214–33.

Malhotra, N. "Completion Time and Response Order Effects in Web Surveys." *Public Opinion Quarterly* 72, no. 5 (December 15, 2008): 914–34.

Teitcher, J.E., W.O. Bockting, J.A. Bauermeister, C.J. Hoefer, M.H. Miner, and R.L. Klitzman. "Detecting, Preventing, and Responding to Fraudsters in Internet Research: Ethics and Tradeoffs." *Journal of Law, Medicine and Ethics* 43, no. 1 (March 15, 2015): 114–31.

Tourangeau, R., F.G. Conrad, and M.P. Couper. *The Science of Web Surveys.* Oxford: Oxford University Press, 2013.

Tourangeau, R., M.P. Couper & C. Frederick. "Spacing, Position, and Order: Interpretive Heuristics for Visual Features of Survey Questions." *Public Opinion Quarterly* 68, no. 3 (October 1, 2004): 368–93.

Walker, R., R. Pettit, and J. Rubinson. The Foundations of Quality Study Executive Summary 1: Overlap, Duplication, and Multi Panel Membership. New York: The Advertising Research Foundation. 2009.

Yan, T., and R. Tourangeau. "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times." *Applied Cognitive Psychology* 22, no. 1 (January 1, 2008): 51–68.

Zhang, C., C. Antoun, H.Y. Yan and F.G. Conrad. "Professional Respondents in Opt-in Online Panels: What Do We Really Know?" *Social Science Computer Review* 38, no. 6 (December 1, 2020): 703–19.

ACKNOWLEDGEMENTS

We would like to thank J. Michael Dennis for his review of this brief.

ABOUT NORC

NORC at the University of Chicago conducts research and analysis that decision-makers trust. As a nonpartisan research organization and a pioneer in measuring and understanding the world, we have studied almost every aspect of the human experience and every major news event for more than eight decades. Today, we partner with government, corporate, and nonprofit clients around the world to provide the objectivity and expertise necessary to inform the critical decisions facing society.