2. Diagnosing survey response quality

D. Sunshine Hillygus and Tina LaChapelle

Getting quality survey responses is an increasingly difficult task for public opinion research. A growing literature has documented concerns about low-quality respondents, particularly in non-probability samples (e.g., Bowling & Huang, 2018). As one practitioner wrote, "Efforts to reduce the number of poorly engaged respondents have become an industry obsession" (Gittelman & Trimarchi, 2012, p. 1). Estimates about the prevalence of data quality issues due to carelessness or fraudulent behavior vary greatly between studies, methods, and diagnostic metrics used, but even a seemingly small number of careless responses can impact survey results (Ahler, Roush, & Sood, 2019; Huang, Liu, & Bowling, 2015; C. Kennedy et al., 2020). Inattentive or careless respondents can jeopardize the reliability and validity of survey estimates, and—ultimately—any knowledge claims being made from surveys.

Researchers have relied on a variety of approaches to mitigate these threats, such as attention check or trap questions and post hoc speeding checks. Some vendors have even created real-time methods to identify and remove inattentive or fraudulent respondents (Cooke & Regan, 2008). Currently, however, efforts to identify careless respondents are often ad hoc and haphazard. The field lacks a clear framework or standard practices for diagnosing and handling response quality issues. This chapter reviews various data quality metrics, including their strengths and limitations, and outlines a future research agenda needed to advance the field.

DEFINING DATA QUALITY

In recent years, any mention of survey results almost inevitably turns to data quality concerns, but there is wide variation in the exact nature of such concerns. For example, there have been extensive debates about the source of polling errors in the 2020 United States presidential election, which saw the worst polling performance in decades (Clinton et al., 2021). Data quality discussions were also prompted by the discovery of respondents faking their location or using "bots" to submit multiple surveys on Amazon MTurk (Ryan, 2018). Likewise, a viral TikTok video created massive gender and age skews in samples of Prolific users, affecting the demographic balance of researchers' surveys (Charalambides, 2021). These examples represent very different data quality issues, but all raise concerns about the extent to which the resulting data would be "fit for purpose"—the simplest definition of data quality (Biemer & Lyberg, 2003).[1]

For surveys to provide a valid estimate of the population of interest requires both a representative sample and accurate survey responses. In this chapter, we focus on evaluating threats to survey response accuracy from inattentive or fraudulent respondents. To be clear, this type of measurement error is just one way that recorded survey responses can deviate from their true values, and measurement error is just one aspect of overall data quality.[2] A key challenge for the field is that the existing theoretical framework for assessing survey data quality is not well suited to the non-probability surveys that increasingly dominate political science research.

For probability samples, the Total Survey Error (TSE) paradigm (e.g., Groves & Lyberg, 2010) offers a conceptual framework for assessing the different sources of error that may arise in the design, collection, processing, and analysis of survey data. This can be broadly divided into errors that impact the representativeness of the sample and errors that impact the accuracy of survey measurements. However, the errors that impact representativeness of a probability-based survey sample (coverage error, sampling error, and non-response error) do not generally apply to non-probability samples—coverage and non-response errors are typically unknown and sample generalizability rests on modeling assumptions that do not easily fit within the TSE framework (Baker et al., 2010). A rich empirical literature has evaluated the extent to which various non-probability samples appear representative of a broader population with somewhat mixed results; some non-probability surveys can produce results comparable to probability samples (e.g., Ansolabehere & Schaffner, 2014), while others show significant deviations from known benchmarks (e.g., Callegaro et al., 2014). In recognition of concerns about representativeness, many researchers consider non-probability samples to be "fit for purpose" for randomized survey experiments, even if they do not consider them to be of sufficient quality for producing population estimates (Baker et al., 2010).[3] Regardless, researchers must still explicitly evaluate the quality of survey responses given widespread concern that respondents may provide erroneous data due to carelessness, confusion, or dishonesty.

To be sure, response quality is a concern for all surveys, but it is especially pronounced for online surveys. It is well documented that survey respondents reduce the effort they invest in answering questions when there is no interviewer to prompt, probe, and follow up (Heerwegh & Loosveldt, 2008). If respondents fail to fully engage in the cognitive process necessary to answer a survey question, the integrity of survey responses can be impacted. Krosnick (1991) calls this survey satisficing. Survey satisficing occurs when respondents fail to carefully and thoroughly perform all the cognitive steps required to answer a survey question: (1) comprehending the question; (2) retrieving relevant information; (3) integrating this information into a required judgment; and (4) selecting and reporting the appropriate answer (Tourangeau, Rips, & Rasinski, 2000). It is also the case that respondents sometimes provide *deliberately* inaccurate responses as expressive responding (e.g., partisan cheerleading), due to social desirability, to maximize incentives, or just as a malicious attempt to corrupt the data. As one example of fraudulent responding, MTurk samples restricted to United States participants routinely find responses associated with non-United States IP addresses or taken from virtual private networks that mask the IP address and geolocation (R. Kennedy et al., 2020).

Online surveys are also particularly at risk for response quality issues because they are often recruited from opt-in panels or other platforms where people sign themselves up to get money or other rewards for taking surveys. Respondents who are taking a lot of surveys might be inattentive to any one survey, and thus more likely to engage in satisficing. Some respondents might also answer in a fraudulent way in order to qualify for higher incentives or more surveys (Devine et al., 2013). One study found that 14 percent of survey participants claimed to own a Segway human transporter, despite Segway ownership being exceedingly rare (Downes-Le Guin, Mechling, & Baker, 2006). Another recent study attempting to sample current and former members of the U.S. Army from online nonprobability panels found that 81% of respondents misrepresented their military credentials (Bell and Gift, 2022).

Importantly, this type of measurement error does not simply add random noise. If inattention just increased random error, the statistical estimates would exhibit higher variance, but remain unbiased. Unfortunately, research has shown that fraudulent and careless responding

can also bias policy-relevant estimates, making inattention a more concerning threat to survey results (Malone & Lusk, 2018). Such inaccuracies can lead to incorrect conclusions about associations between variables (e.g., Bernstein, Chadha, & Montjoy, 2001; Dahlgaard, Hansen, Hansen, & Bhatti, 2019) and incorrect assumptions about population-level estimates (e.g., Bullock & Lenz, 2019). Thus, it is increasingly important for researchers to explicitly evaluate survey response quality.

METRICS OF SURVEY RESPONSE QUALITY

How can we diagnose data quality issues associated with respondents providing inaccurate or fraudulent responses? Before turning to specific metrics that can be used for evaluating response quality, three key take-home messages are worth keeping in mind. First, the best way to address data quality concerns is to avoid them in the first place. Researchers can ensure high-quality responses by designing surveys that follow best practices for question wording, question order, questionnaire visualization, pretesting, and the like.[4] At the same time, researchers need to have a clear understanding of response quality and an assessment plan prior to the start of data collection because it will influence the study design and data management. That means, for instance, programming the survey to collect the appropriate paradata, such as response times, and including the necessary questions to identify careless or fraudulent respondents.

Second, there is no one silver bullet for identifying "bad" respondents. Response quality metrics are imperfect tools that themselves have measurement error (Maniaci & Rogge, 2014; Thomas & Clifford, 2017). Any one diagnostic can miss some inattentive respondents or can inappropriately flag some otherwise valid cases. Even "good" respondents can skim or misunderstand directions or accidently select the wrong response options.

Third, deleting "bad" cases can do more harm than good. At minimum, throwing out cases reduces statistical power and is akin to throwing away money. More concerning is that deleting respondents can jeopardize the survey's external validity and study findings by introducing systemic bias (Berinsky, Margolis, & Sances 2014; Hillygus, Jackson, & Young, 2014; Lancsar & Louviere, 2006). For example, previous research has shown that respondents who fail attention checks are more likely to be younger, male, less educated, and non-white (Berinsky et al., 2014; Downs, Holbrook, Sheng, & Cranor, 2010; Maniaci & Rogge, 2014; Ramsey, Thompson, McKenzie, & Rosenbaum, 2016). Omitting inattentive respondents can bias associations between variables (Bernstein et al., 2001; Dahlgaard et al., 2019)—sometimes attenuating and sometimes inflating observed correlations (see Huang et al., 2015). In experimental studies, dropping respondents can induce asymmetry across treatment arms and introduce bias in the results (Aronow, Baron, & Pinson, 2019).

Although more research is needed, the emerging consensus is that it is preferable to use multiple response quality flags and then to check the sensitivity of survey findings (Geisen, Smith, & Belden, 2021; Phillips, 2015; Thomas & Clifford, 2017). The quality metrics we outline below—attention checks, speeding, straightlining, item non-response, open-ended quality checks, and self-reported measures—can be used as flags, with a summary measure indicating how many quality flags an individual failed. Individuals who fail more than one flag are far more likely to provide consistently low-quality responses (Geisen et al., 2021; Phillips, 2015). Using these flags, researchers should then check the sensitivity of their results to response

quality issues, reporting both results and the assessment process. Transparent reporting and documentation of the metrics used is foundational to ensuring data quality.[5]

ATTENTION CHECKS

In an effort to identify problematic participants, many researchers include attention checks or "trap questions" to catch inattentive respondents (e.g., Berinsky, Margolis, Sances, & Warshaw, 2021; Liu & Wronski, 2018). Trap questions are intended to identify respondents who fail to read the survey instructions, who randomly respond to a survey item, or who might be providing an inaccurate response (e.g., in an effort to qualify for an incentive). Although the use of attention checks is increasingly common for survey research, there is considerable variability in exactly how they are operationalized and used. Some survey vendors use trap questions to disqualify a respondent (Oppenheimer, Meyvis, & Davidenko, 2009), while other researchers use it to train a respondent (Berinsky, Margolis, & Sances, 2016), or use it as one quality metric in a broader assessment/sensitivity analysis (Geisen et al., 2021).

Attention checks come in several different forms. One common type of attention check is called an instructional manipulation check (IMC), where respondents are given a survey question but told in the instructions to ignore the provided response options, giving proof they have read the question some other way (Anduiza & Galais, 2017; Berinsky et al., 2014; Oppenheimer et al., 2009). For example, Oppenheimer et al. (2009) propose the blue-dot task, in which respondents are instructed to ignore the Likert scale response options arranged from "very rarely" to "very frequently," and instead click a little blue dot at the bottom of the screen. Relatedly, instructed response items are a type of IMC that instructs respondents to mark a specific response category (e.g., "click strongly agree") regardless of how a respondent might otherwise answer the question. Another type of attention check, the bogus item, provides a standard set of response options, but is a question for which there is only one correct answer. For example, in a survey taken by undergraduate psychology students, Meade and Craig (2012) included in a list of 50 agree–disagree statements the item "I am currently enrolled in a psychology course." If the respondent selects the incorrect answer to an unambiguously true statement, it is a clear indication that they are responding randomly or carelessly.

Although attention checks are popular, recent research highlights that they can have downsides. Attention checks are themselves subject to measurement error—otherwise good survey respondents make errors, so a respondent who fails an attention check at one point in the survey will not necessarily fail other ones (Thomas & Clifford, 2017). Research has shown, for instance, wide variation in the percentage of respondents who fail a given attention check question depending on the design of the question and location within the survey (Phillips, 2015). IMCs and instructed response items may show a lack of trust in the respondent, so they can contribute to a poor user experience with the survey. Respondents can find trap questions misleading or irritating, potentially increasing other types of bad behavior, such as survey break-offs (Vannette, 2017). Attention check questions can also induce changes in respondent behavior. Hauser and Schwarz (2015) show that attention checks are not just measures of attention, but are also interventions that spark respondents to have a more deliberative mindset in answering subsequent questions, potentially impacting study results. They can also induce socially desirable responding, where respondents edit, adjust, or censor their answers to fit what is socially acceptable and expected because the attention checks create a sense of being

watched (Clifford & Jerit, 2015). Those who fail attention checks are not random, so deleting cases can increase demographic biases in the sample (e.g., Anduiza & Galais, 2017; Berinsky et al., 2014), which is especially problematic if the question is located post-treatment.

Equally problematic are attention checks that try to trick respondents through sneaky, confusing, or nonsensical questions. The survey vendor Prolific emphasizes that attention checks are fair only if they are checking whether a participant pays attention to the question rather than the question instructions.[6] For example, the following attention check question would be unfair because a respondent would be flagged as inattentive even if they answered the question correctly: "What color is grass? The fresh, uncut grass, not leaves or hay. Make sure to select purple as an answer so that we know you are paying attention." Moreover, Curran and Hauser (2019) show that using silly or nonsensical items does not allow for a distinction between inattention, overthinking, and mischievous responding. For example, in explaining why they had selected affirmatively to the bogus item "I am paid bi-weekly by leprechauns," the respondent explained "I am paid bi-weekly, just not by leprechauns."[7] Curran and Hauser (2019) recommend including bogus items that are based on simple known truths—"I live outside the United States" or "I have never used a computer" are two of our go-to attention checks for online surveys.

Including attention checks in online surveys has become the norm, but it is necessary to think through the inherent trade-offs associated with using them. Attention checks may not only make respondents feel frustrated but also induce changes in respondent behavior that fundamentally affect a study's findings. For this reason, we recommend that attention checks—should they be used—be kept fair and simple, and be treated as a quality flag rather than a basis for exclusion, especially if located deep in a survey. It is also worth noting that probability samples find far lower rates of attention check failures (C. Kennedy et al., 2020), suggesting their inclusion might not be worth the potential trade-off for such designs. There are alternative metrics, such as those reviewed below, that researchers can use to assess the quality of their survey data.

SPEEDING

Fast responding (or "speeding") is one of the most commonly used metrics of satisficing and low-quality data. Short response times can indicate a lack of attention on the part of respondents. Respondents who give answers quickly are assumed to have not given much thought to the answers and suggest that respondents might be looking to finish the questionnaire as fast as possible rather than to provide careful and accurate responses (Callegaro et al., 2014; Greszki, Meyer, & Schoen, 2015; Malhotra, 2008).

Conceptually, speeding is straightforward. The challenge is determining how fast is too fast in practice. Respondents can sometimes answer a question both quickly and accurately depending on factors such as task difficulty and accessibility of an attitude. The optimal response time depends on the reading skills of the respondent—which vary widely across the population—and the survey questions being asked. Open-ended questions following long vignettes obviously take longer than a short yes/no question. Poorly written survey questions (e.g., jargony language, double-barreled questions) also increase cognitive effort, increasing response time (Bassili & Scott, 1996). Research has shown that frequent survey takers, like

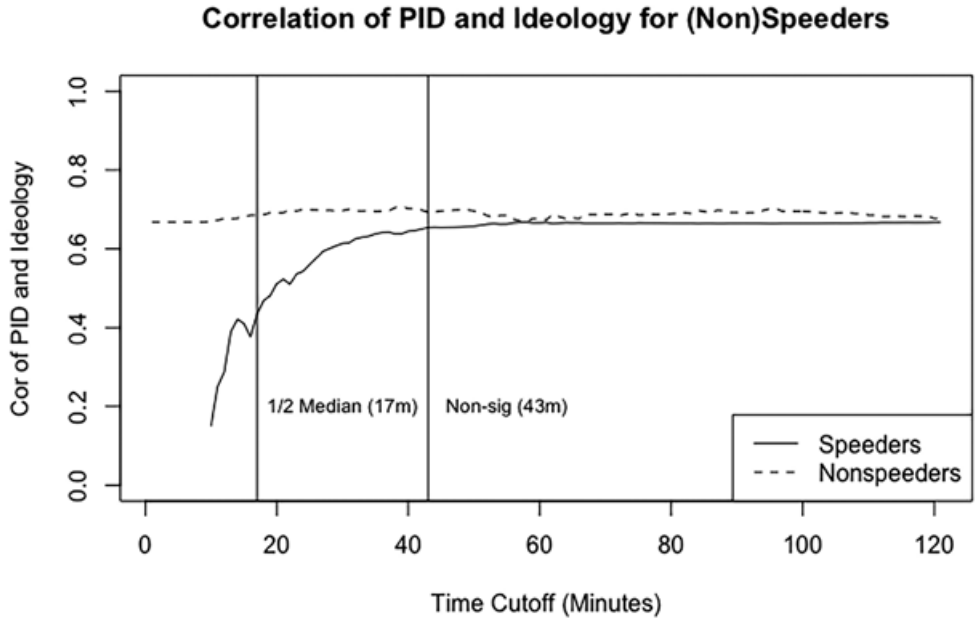
panelists in an opt-in non-probability panel, answer more quickly than fresh recruits without impacting response quality (Hillygus et al., 2014).

The particular response time metric used in previous research also varies widely in practice.[8] Response time can be measured with respect to a question, set of questions, or the entire questionnaire.[9] The definition of speeding also varies—some use an external threshold, like average reading speed, while others look for statistical outliers in the observed data. As an example of the former, Zhang, Antoun, Yan, and Conrad (2020) set the speeding threshold at 300 milliseconds per word based on the logic that the average reading speed among college students for comprehension is about 200 milliseconds per word.[10] Other researchers look for response time outliers in the collected data; for example, they might exclude those who took less than 5 minutes on a survey. Often, the selected threshold is arbitrary. Some researchers identify outliers based on a statistical threshold of the collected data, with wide variation in the particular threshold used—1st/99th percentile, 5th/95th percentile, or +/− 2 standard deviations from the mean response time (Christian, Parsons, & Dillman, 2009). Because online surveys can also have respondents with very slow response times, it is generally preferable to rely on median, rather than mean, in identifying outliers. One rule of thumb used by industry practitioners is flagging anything faster than half the median response time, with some recommending calculating duration excluding the last few questions in the survey in case professionalized respondents wait on the last question to avoid being flagged for speeding. It is also important that any speeding metrics account for variation in the number or length of survey questions across respondents (e.g., due to branching logic or treatment conditions). In such cases, speeding measures should be restricted to common survey items or conditional on the treatment.

Although it is increasingly common to use a speeding metric to identify and remove respondents, more research is needed both to define speeding thresholds and to evaluate when removal is appropriate. That is, when does the risk of deleting valid respondents outweigh the risk of deleting too few inattentive respondents? Some research suggests that removing "too fast" responses does not alter marginal distributions and has negligible impact on coefficient differences (Greszki et al., 2015). At the most extreme levels of speeding, however, there are clear impacts on expected correlations, as can be seen in our analysis in Figure 2.1 from the 2019 American National Election Study (ANES) pilot survey, a YouGov sample.[11] The figure maps the correlation between party identification and ideology across different speeding thresholds. The first vertical line represents the half median threshold, the second vertical line represents the point at which the differences in correlation between the speeders and non-speeders are not statistically different from one another. Thus, the half median threshold appears to flag problematic respondents (9.6 percent of the sample); at the same time, it is also clear that there is no one threshold that cleanly separates "bad" and "good" respondents. A conservative approach would be to omit only those with an excessively fast speed (e.g., quarter median), then to create a response quality flag for those who were faster than half median.

There clearly remain many opportunities for further research on using response time for improving response quality. For one, most researchers focus only on speeders, whereas we know much less about the quality impacts from excessively slow responding (e.g., Christian et al., 2009). There is also a question of whether or not it might be possible to train respondents; that is, get them to slow down while they are taking the survey, rather than just diagnosing speeders after the fact. Previous research has found mixed success with interventions to slow down respondents (e.g., Conrad, Tourangeau, Couper, & Zhang, 2017; Zhang & Conrad, 2018). Generally, such interventions can reduce speeding, but they can have trade-offs—

they may also annoy respondents, triggering higher rates of break-offs (Conrad et al., 2017; DeRouvray & Couper, 2002).



Note: 2019 American National Election Study pilot. Median completion time was 34 minutes. Data are unweighted (n = 3165). Vertical lines represent the half median and the point at which speeders and non-speeders are no longer statistically different.

Figure 2.1 Correlation of party and ideology by speeding threshold

STRAIGHTLINING

Another potential indicator of survey satisficing is non-differentiation or "straightlining"—when a respondent provides identical responses to multiple items with the same scale, such as responding "agree strongly" to back-to-back items in a series (Herzog & Bachman, 1981). Straightlining is more prevalent in grid or matrix questions in self-administered surveys (Roßmann, Gummer, & Silber, 2018; Tourangeau, Couper, & Conrad, 2004). For this reason, many survey methodologists advise against the use of long series of survey grids when designing a questionnaire (Dillman, Smyth, & Christian, 2014). Nonetheless, they remain a prominent question format and, when designed appropriately, can serve as a data quality metric.

Although many survey practitioners use straightlining as one, and sometimes the only, indicator for respondent inattention (e.g., Fricker, Galesic, Tourangeau, & Yan, 2005), it is critical that it has been designed in a way that straightlining is a meaningful quality metric. Straightlining often represents a valid response pattern (Reuning & Plutzer, 2020). For example, many psychological scales include a series of items all with the same directional valence (e.g., life satisfaction), so that providing the same response to every item is a valid

response option rather than an indication that the respondent has not given much thought to the answer.

For straightlining to be a metric of response quality, it requires inclusion of negatively correlated items in the series. Reuning and Plutzer (2020) show that including a reverse-coded item reduces 98 percent of valid straightlining. Reverse-coded items, however, must be designed carefully. It is important to avoid double negatives and confusing phrasing (e.g., Johnson, Bristow, & Schneider, 2004). Respondents can understandably overlook a reverse-coded item if it comes deep in a long list of items. We recommend that a reverse-coded item be located early in a grid—it should be thought of as a speed bump that serves as a reminder to respondents to read carefully. Another design consideration is the inclusion of a middle category, which often creates a valid straightline response even if a reverse-coded item is included in the grid. Ideally, then, a straightlining metric is used with a grid that includes at least one reverse-coded item, located early in the grid, and that excludes a middle category.

ITEM NON-RESPONSE

Item non-response—the failure to provide an answer to an individual survey question—can be another indication of respondent satisficing if respondents provide "no opinion" rather than going through the cognitive effort of generating a valid response. Research has found that item non-response is more common in self-administered surveys, in long questionnaires, and on higher-burden questions (Baker et al., 2010). A recent meta-analysis of 141 satisficing publications found that item non-response measures were the most commonly used indicator of satisficing (Roberts, Gilbert, Allum, & Eisner, 2019). Item non-response can take the form of "item refusals" (REF) or "don't know" (DK) responses,[12] and quality metrics can be calculated as a percentage of no-opinion responses for the entire questionnaire or some subset of questions.

As with other data quality metrics, item non-response is not a perfect indication of a "bad" respondent. There are a variety of reasons why someone might give no opinion. No-opinion responses can reflect not only a lack of motivation to give a valid answer, but also an inability to do so. Previous research finds that item non-response is more common for questions that are difficult to understand, require arduous memory recall, or ask for complex mental calculations (Bishop, Tuchfarber, & Oldendick, 1986). Item non-response can also be a response to a poorly worded or confusing question (Schuman & Presser, 1996). Interviewer characteristics and experience can impact how willing respondents are to answer (Holbrook, Green, & Krosnick, 2003). Finally, item non-response is more common for sensitive questions—respondents may carefully read a survey question, but still be unwilling to respond to it because of concerns about confidentiality (Olson, Smyth, & Ganshert, 2019).

Survey mode is a further consideration in using item non-response as a quality metric. DK responses tend to be volunteered responses in interviewer-administered surveys—accepted by the interviewer but not an explicit response option. In contrast, it is typically recommended that online questionnaires exclude the DK options precisely because it can be viewed as an invitation to avoid formulating a thoughtful response (Vannette & Krosnick, 2017). Our rule of thumb for online questionnaire design is to offer a DK response only when it can be analyzed as a substantive response—for example, with knowledge questions. When calculating

a quality flag, it is important to exclude any questions where DK could be interpreted as a substantively meaningful response.

Increasingly, surveys using online panels or crowdsourcing platforms have so little item non-response as to not be a terribly useful metric. Some survey vendors "clean" the data by omitting respondents who fail to respond to a certain percentage of questions or imputing responses for some items left blank.[13] Some vendors and/or researchers discourage skipping questions through the use of follow-up prompts, forced choice questions, or threats to reject incentive payments. Even in cases where a researcher might allow a respondent to skip a question, frequent survey takers may have become reluctant to do so through previous experience. Empirical research indeed finds that experienced panelists are less likely to give a no-opinion response (Smith & Brown, 2006; Binswanger, Schunk, & Toepoel, 2013). For example, a recent Pew study found that no more than 2 percent of respondents from an opt-in panel gave a blank, DK, or REF response, compared to 13–19 percent of freshly recruited respondents from an address-based sample (C. Kennedy et al., 2020, p. 13).

Given the limitations of calculating and interpreting item non-response measures, we find it more straightforward to include other metrics of engagement in the questionnaire, such as an open-ended question as described below.

OPEN-ENDED QUESTIONS

Although closed-ended questions, in which predefined response options are provided, are most common in surveys, open-ended questions offer one of the clearest indications of data quality. Since the origins of surveys as a field, researchers have debated the merits and limitations of open- versus closed-ended questions (e.g., Schuman & Presser, 1979). Open-ended questions help collect deeper insights, but also require greater cognitive burden and induce survey fatigue in respondents. It is well documented that they are more burdensome for the researcher to analyze and more burdensome for the respondent to answer (Dillman et al., 2014). At the same time, low-effort responses to a well-designed open-ended question can be easier to spot and offer high face validity compared to closed-ended questions. A lack of engagement is clear if respondents skip the open-ended question or provide an irrelevant or inadequate answer.

Because some experienced respondents will avoid skipping items, the content of a response can also reveal a lack of engagement. Some researchers rely on word count in open-ended questions as an indicator of satisficing (Callegaro et al., 2014). Others look for indications of gibberish or non-sequitur answers (Hillygus et al., 2014). In a recent comprehensive study (C. Kennedy et al., 2020), Pew identifies as satisficing the following kinds of non-sequitur answers: unsolicited product reviews, plagiarized text from other websites found when entering the question in a search engine, conversational text, common words not matching the question, gibberish, or a no-opinion response (meaning the respondent either left the box blank or gave a DK or REF answer). While most respondents gave valid answers, the study found that 2–4 percent of opt-in poll respondents provided junk answers. These responses all offer a clear indication of survey satisficing, if the open-ended question is appropriately designed.

In selecting an open-ended question to include on a survey, it should be one that is easy to answer and analyze (e.g., avoiding knowledge or recall questions). Some examples include the standard most important issue question, "What do you think is the most important problem facing the country today?" or, as Pew asked in their recent study, "What would you like to

see elected leaders in Washington get done during the next few years?" Although open-ended responses are typically more burdensome to code and analyze than closed-ended responses, we have found that coding as a quality metric tends to be fast and straightforward, even if the question responses are not otherwise used.

SELF-REPORT MEASURES

A final technique for assessing response quality is to directly ask respondents about it. Such questions ask self-reported levels of attention, speed, or honesty in responding to the survey. Although these items are easily detected as "quality checks" by the respondent and can themselves be prone to dishonest answers, research finds they can improve data validity (Aust, Diedenhofen, & Ullrich, 2013).

Self-reported honesty assessments have long been used in surveys about sensitive topics, such as sexual behavior and drug use (e.g., Brown et al., 2012). For example, the Rochester AIDS Prevention Project for Youth asked, "Overall, how honest were you in answering this questionnaire?" with the seven-point response options ranging from "completely dishonest" to "completely honest" (Siegel, Aten, & Roghmann, 1998).[14] Similar direct measures have been used to capture inattentive responding. For example, Ward, Meade, Allred, Pappalardo, and Stoughton (2017) asked "I gave this study ___ attention": "almost no," "very little of my," "some of my," "most of my," "my full." Some researchers simply ask "In your honest opinion, should we use your data?" at the end of a questionnaire (Meade & Craig, 2012). In one of the most comprehensive studies, Maniaci and Rogge (2014) developed a psychometric scale of inattention, evaluating 67 potential items, with subscales for careless responding, patterned responding, rushed responding, and instruction skipping.[15] In an effort to reassure incentive-motivated respondents, these questions tend to be the final questions of the survey and sometimes provide a reassurance like "Your response to this question will have no impact on your compensation."

Self-reported measures can be used to capture those who are inattentive as well as those who might give insincere responses intentionally for the sake of being provocative, inflammatory, or humorous—also called mischievous responding or survey trolling (Lopez & Hillygus, 2018). For example, in-person follow-up interviews of the high-quality Adolescent Health Survey (Add Health) found that 99 percent of respondents who had reported having artificial limbs in the self-complete portion of the survey had not given an accurate response (Fan et al., 2006).[16] The following question added at the end of the survey can help to identify such respondents: "We sometimes find people don't always take surveys seriously, instead providing funny or insincere answers. How often did you give a serious response to the questions on this survey? Never serious, Some of the time serious, About half of the time serious, Most of the time serious, Always serious." Lopez and Hillygus (2018) find that individuals who self-report answering questions insincerely are far more likely to claim to believe in a conspiracy theory. These individuals are not, however, the same as those who are flagged for inattention. In the 2019 ANES pilot reviewed earlier, the correlation between being flagged as a speeder and being flagged as a troll was just 0.38. A potential advantage of this item over the classic honesty self-reports is that admitting to giving a funny response might carry less stigma than admitting to lying. Although this approach will obviously not catch those who will

lie about lying, a non-trivial number of respondents will admit that their responses are of poor quality, which offers one more quality check for the survey data (Curran, 2016).

CONCLUSION

As online non-probability surveys become more commonplace in public opinion research, there are growing concerns about response quality issues in the resulting data. Online respondents may often provide responses that are inaccurate—due to either carelessness and inattention, reluctance to engage with difficult or burdensome questions, or mischievous intentions. Detecting quality issues through multiple checks is paramount for researchers relying on online surveys to build our understanding of public opinion, knowledge, or behavior.

The metrics previewed here offer a partial and incomplete set of approaches for detecting low-quality respondents. The proposed metrics will not always be appropriate for all survey projects, and the researchers may not be willing to surrender valuable survey real estate to all of the possible quality metrics. Nonetheless, researchers should consciously and thoughtfully plan an explicit response quality assessment. To summarize our recommendations:

1. All surveys, regardless of vendor or sampling design, should screen for careless or fraudulent respondents.
2. Researchers should develop a plan to evaluate response quality *in advance of data collection* as many data quality metrics require design into the questionnaire or data collection process. Ideally, this plan would be documented—as part of a pre-registration plan, for instance.
3. Multiple detection methods should be used to identify problematic respondents. No one detection method should be used in isolation.
4. If respondents are excluded, exclusion criteria should be conservative. The effects of exclusion should be analyzed and reported.
5. Above all else, transparency is the key to assessing data quality. Errors are inevitable. It is only possible to evaluate their implications with sufficient information about the data collection process.

Although it is a nascent field of study, the existing research highlights the need to take seriously survey response quality in online non-probability samples. Surveys that rely on crowdsourced platforms or online panels show widespread evidence of careless and fraudulent responses. Furthermore, research shows that relying on such error-prone data will result in biased or wrong conclusions if we do not address the quality issues and consequent problems. At the same time, the field must recognize that it is not enough to simply classify data as tin or gold. Rather, we need to be able to extract value from imperfect data—through the development of new metrics for diagnosing quality issues, new approaches for evaluating the sensitivity of results to quality issues, and the development of guidelines and standards. More research is needed to help develop standards for more precisely identifying inattention and for separating out poor respondent engagement from poor questionnaire design. Finally, the acknowledgment and documentation of errors in survey data should be routinized to demonstrate scientific integrity and ensure the credibility of our substantive findings.

NOTES

1. A consensus has emerged among national and international statistical offices that data quality is a multidimensional concept that can be broadly defined as "fitness for use"—that is, survey data need to be accurate enough to achieve their intended purpose, be available at the time needed (timely), and accessible to those for whom the survey was conducted (Biemer & Lyberg, 2003).
2. Other sources of measurement error include the survey instrument—e.g., due to question wording, order, or context—and the interviewer.
3. There remains considerable debate about validity and reliability of experimental results using non-probability samples (e.g., Berinsky, Huber, & Lenz, 2012; Chandler & Paolacci, 2017).
4. See Dillman et al. (2014) and Groves et al. (2011) for recommended practices in survey methodology.
5. For survey methodology disclosure guidelines, see the reporting standards for the American Association of Public Opinion Research or the Quality Reports proposed by Eurostat. www.aapor.org/Standards-Ethics/AAPOR-Code-of-Ethics.aspx and https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf/bf98fd32-f17c-31e2-8c7f-ad41eca91783?t=1583397712000.
6. See Prolific's guide on attention checks here: https://researcher-help.prolific.co/hc/en-gb/articles/360009223553-Using-attention-checks-as-a-measure-of-data-quality.
7. Other problematic bogus items used in previous research can be unknown truths ("All my friends say I would make a great poodle"), semantic argument truths ("I can teleport across time and space"), sliding-scale truths ("I can run 2 miles in 2 mins"), and double-barreled truths ("I am paid bi-weekly by leprechauns").
8. In a systematic literature review, Christian et al. (2009) identified 28 papers published between 2003 and 2017 examining response times: 20 excluded outliers (an average of 4 percent of the sample) but rarely reported how exclusion impacted results.
9. We also encourage inclusion of a question-specific timer for any long questions (e.g., vignettes), knowledge questions, or open-ended questions.
10. This threshold has been shown to be related to straightlining—another commonly used indicator for careless responding (e.g., Zhang & Conrad, 2014).
11. The 2019 ANES pilot was fielded by a YouGov survey firm, an opt-in non-probability panel, in December 2019. The target population is United States citizens at least 18 years of age. For full methodology, including question wording, see https://electionstudies.org/wp-content/uploads/2020/02/anes_pilot_2019_userguidecodebook.pdf.
12. Some argue they reflect different underlying processes; DK is more likely if information is not accessible, while a REF answer is more likely due to social desirability (Olson et al., 2019).
13. Although it is not always apparent to the user, imputation of demographic and political profile characteristics is common practice for many online panels.
14. Others use honesty or seriousness pledges *prior* to asking questions; e.g., McDonald, Scott, and Hanmer (2017) find an honesty pledge reduces turnout overreporting by 11 percentage points.
15. See Brühlmann, Petralito, Aeschbach, and Opwis (2020) for a comparison of these items to other measures, such as speeding and attention checks.
16. Mischievous respondents can also be diagnosed through analysis of low-incident characteristics (Robinson-Cimpian, 2014).

REFERENCES

Ahler, D. J., Roush, C. E., & Sood, G. (2019, April). *The micro-task market for lemons: Data quality on Amazon's Mechanical Turk*. Paper presentation, 77th Annual Conference of the Midwest Political Science Association, Chicago, IL.

Anduiza, E., & Galais, C. (2017). Answering without reading: IMCs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519.

Ansolabehere, S., & Schaffner, B. F. (2014). Does survey mode still matter? Findings from a 2010 multi-mode comparison. *Political Analysis*, *22*(3), 285–303.
Aronow, P. M., Baron, J., & Pinson, L. (2019). A note on dropping experimental subjects who fail a manipulation check. *Political Analysis*, *27*(4), 572–589.
Aust, F., Diedenhofen, B., & Ullrich, S. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535.
Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M. et al. (2010). Research synthesis: AAPOR report on online panels. *Public Opinion Quarterly*, *74*(4), 711–781.
Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, *60*(3), 390–399.
Bell, A. M., & Gift, T. (2021). Fraud in online surveys: Evidence from a nonprobability, subpopulation sample. *Journal of Experimental Political Science*, 1–6.
Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368.
Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739–753.
Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers? *Journal of Experimental Social Psychology*, *66*, 20–28.
Berinsky, A. J., Margolis, M. F., Sances, M. W., & Warshaw, C. (2021). Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods*, *9*(2), 430–437.
Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting, voting: Why it happens and why it matters. *Public Opinion Quarterly*, *65*(1), 22–44.
Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*, Vol. 335. John Wiley & Sons.
Binswanger, J., Schunk, D., & Toepoel, V. (2013). Panel conditioning in difficult attitudinal questions. *Public Opinion Quarterly*, *77*(3), 783–797.
Bishop, G. F., Tuchfarber, A. J., & Oldendick, R. W. (1986). Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly*, *50*(2), 240–250.
Bowling, N. A., & Huang, J. L. (2018). Your attention please! Toward a better understanding of research participant carelessness. *Applied Psychology*, *67*(2), 227–230.
Brown, J. L., Sales, J. M., DiClemente, R. J., Salazar, L. F., Vanable, P. A., Carey, M. P., Brown, L.K., Romer, D., Valois, R.F., & Stanton, B. (2012). Predicting discordance between self-reports of sexual behavior and incident sexually transmitted infections with African American female adolescents: Results from a 4-city study. *AIDS and Behavior*, *16*(6), 1491–1500.
Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, *2*, 100022.
Bullock, J. G., & Lenz, G. (2019). Partisan bias in surveys. *Annual Review of Political Science*, *22*, 325–342.
Callegaro, M., Baker, R. P., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (Eds) (2014). *Online panel research: A data quality perspective*. John Wiley & Sons.
Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, *8*(5), 500–508.
Charalambides, N. (2021). We recently went viral on TikTok—here's what we learned. Blog post, *Prolific.* https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned/
Christian, L. M., Parsons, N. L., & Dillman, D. A. (2009). Designing scalar questions for web surveys. *Sociological Methods and Research*, *37*(3), 393–425.
Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly*, *79*(3), 790–802.
Clinton, J. D., Agiesta, J., Brenan, M., Burge, C., Connelly, M., Edwards-Levy, A. et al. (2021). *Task force on 2020 pre-election polling: An evaluation of the 2020 general election polls*. American Association for Public Opinion Research.
Cooke, M., & Regan, S. (2008). ASC Conference—Orion: Identifying inattentive or fraudulent respondents. *International Journal of Market Research*, *50*(5), 695–698.

Conrad, F., Tourangeau, R., Couper, M., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, *11*(1), 45–61.
Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.
Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality*, *82*, 103849.
Dahlgaard, J. O., Hansen, J. H., Hansen, K. M., & Bhatti, Y. (2019). Bias in self-reported voting and how it distorts turnout models: Disentangling nonresponse bias and overreporting among Danish voters. *Political Analysis*, *27*(4), 590–598.
DeRouvray, C., & Couper, M. P. (2002). Designing a strategy for reducing "no opinion" responses in web-based surveys. *Social Science Computer Review*, *20*(1), 3–9.
Devine, E. G., Waters, M. E., Putnam, M., Surprise, C., O'Malley, K., Richambault, C. et al. (2013). Concealment and fabrication by experienced research subjects. *Clinical Trials*, *10*(6), 935–948.
Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method*. John Wiley & Sons.
Downes-Le Guin, T., Mechling, J., & Baker, R. (2006). Great results from ambiguous sources: Cleaning internet panel data. In *ESOMAR World Research Conference: Panel research*.
Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010, April). Are your participants gaming the system? Screening Mechanical Turk workers. *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 2399–2402.
Fan, X., Miller, B. C., Park, K. E., Winward, B. W., Christensen, M., Grotevant, H. D., & Tai, R. H. (2006). An exploratory study about inaccuracy and invalidity in adolescent self-report surveys. *Field Methods*, *18*(3), 223–244.
Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*(3), 370–392.
Geisen, E., Smith, B., & Belden, W. (2021). Evaluation of data quality indicators in online panel providers. Virtual AAPOR Conference.
Gittelman, S. H., & Trimarchi, E. (2012, November). *Rules of engagement: The war against poorly engaged respondents, guidelines for elimination*. Paper presentation, 37th Annual Conference of the Midwest Association for Public Opinion Research, Chicago, IL.
Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the effects of removing "too fast" responses and respondents from web surveys. *Public Opinion Quarterly*, *79*(2), 471–503.
Groves, R. M., Fowler, Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey methodology*, Vol. 561. John Wiley & Sons.
Groves, R. M., & Lyberg, L. (2010). Total survey error: Past, present, and future. *Public Opinion Quarterly*, *74*(5), 849–879.
Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*, *5*(2), 1–6.
Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveying in a high-internet-coverage population: Differences in response quality. *Public Opinion Quarterly*, *72*(5), 836–846.
Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly*, *45*(4), 549–559.
Hillygus, D. S., Jackson, N., & Young, M. (2014). Professional respondents in non-probability online panels. *Online Panel Research: A Data Quality Perspective*, *1*, 219–237.
Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*(1), 79–125.
Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology*, *100*(3), 828–845.
Johnson, J. M., Bristow, D. N., & Schneider, K. C. (2004). Did you not understand the question or not? An investigation of negatively worded questions in survey research. *Journal of Applied Business Research*, *20*(1), 75–86.
Kennedy, C., Hatley, N., Lau, A., Mercer, A., Keeter, S., Ferno, J., & Asare-Marfo, D. (2020). Assessing the risks to online polls from bogus respondents. *Pew Research Center*.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629.
Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.
Lancsar, E., & Louviere, J. (2006). Deleting "irrational" responses from discrete choice experiments: A case of investigating or imposing preferences? *Health Economics*, *15*(8), 797–811.
Liu, M. & Wronski, L. (2018). Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research*, *60*(1), 32–49.
Lopez, J., & Hillygus, D. S. (2018). Why so serious? Survey trolls and misinformation. https://ssrn.com/abstract=3131087 and http://dx.doi.org/10.2139/ssrn.3131087.
Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly*, *72*(5), 914–934.
Malone, T., & Lusk, J. L. (2018). Consequences of participant inattention with an application to carbon taxes for meat products. *Ecological Economics*, *145*, 218–230.
Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83.
McDonald, J. A., Scott, Z. A., & Hanmer, M. J. (2017). Using self-prophecy to combat vote overreporting on public opinion surveys. *Electoral Studies*, *50*, 137–141.
Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437.
Olson, K., Smyth, J. D., & Ganshert, A. (2019). The effects of respondent and question characteristics on respondent answering behaviors in telephone interviews. *Journal of Survey Statistics and Methodology*, *7*(2), 275–308.
Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867–872.
Phillips, K. (2015). *An evaluation of online quality control questions*. Survey Sampling International. Presented at the 2015 Annual Meeting of the American Association of Public Opinion Research.
Ramsey, S. R., Thompson, K. L., McKenzie, M., & Rosenbaum, A. (2016). Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior*, *58*, 354–360.
Reuning, K., & Plutzer, E. (2020, September). Valid vs. invalid straightlining: The complex relationship between straightlining and data quality. *Survey Research Methods*, *14*(5), 439–459.
Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly*, *83*(3), 598–626.
Robinson-Cimpian, J. P. (2014). Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. *Educational Researcher*, *43*(4), 171–185.
Roßmann, J., Gummer, T., & Silber, H. (2018). Mitigating satisficing in cognitively demanding grid questions: Evidence from two web-based experiments. *Journal of Survey Statistics and Methodology*, *6*(3), 376–400.
Ryan, T. J. (2018, August 12). Data contamination on MTurk. Blog post. https://timryan.web.unc.edu/2018/08/12/data-contamination-on-mturk/.
Schuman, H., & Presser, S. (1979). The open and closed question. *American Sociological Review*, 692–712.
Schuman, H., & Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. SAGE Publications.
Siegel, D. M., Aten, M. J., & Roghmann, K. J. (1998). Self-reported honesty among middle and high school students responding to a sexual behavior questionnaire. *Journal of Adolescent Health*, *23*(1), 20–28.
Smith, R., & Brown, H. H. (2006). Data and panel quality: Comparing metrics and assessing claims. *Proceedings of the ESOMAR Panel Research Conference*.
Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197.
Tourangeau, R., Couper, M. P., & Conrad, F. G. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*, 368–393.
Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Vannette, D. L. (2017, June 28). *Using attention checks in your surveys may harm data quality*. Qualtrics. www.qualtrics.com/blog/using-attentionchecks-in-your-surveys-may-harm-data-quality/.
Vannette, D. L., & Krosnick, J. A. (Eds). (2017). *The Palgrave handbook of survey research*. Springer.
Ward, M. K., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, *76*, 417–430.
Zhang, C., Antoun, C., Yan, H. Y., & Conrad, F. G. (2020). Professional respondents in opt-in online panels: What do we really know? *Social Science Computer Review*, *38*(6), 703–719.
Zhang, C., & Conrad, F. G. (2014). Speeding in web surveys: The tendency to answer very fast and its association with straightlining. *Survey Research Methods*, *8*(2), 127–135.
Zhang, C., & Conrad, F. G. (2018). Intervening to reduce satisficing behaviors in web surveys: Evidence from two experiments on how it works. *Social Science Computer Review*, *36*(1), 57–81.