Evaluating the Impact of Review of Open-End Responses for Quality Control

Prepared for AAPOR 2024

Andrea Date, VP Research Methods & Services

Julia Nelson, Sr Research Analyst

The Harris Poll

Background

Open-end review is a common data quality measure

One of the side-effects of incentivizing respondents to take surveys as part of an online panel is that inattentive respondents might be incentivized to provide minimal input and are more prone to things such as recall bias and inconsistent or conflicting responses.

Along with inattentive respondents, there is increasing concern of fraudulent respondents, who intentionally misrepresent their qualifications, to complete and earn survey incentives.

The common belief that inattentive and fraudulent respondents may be impacting data quality necessitate quality control measures.

One of the most time-consuming and costly checks commonly used to assess inattentive and fraudulent respondents is a manual review of the respondent's open-ended responses.

Objectives

Understand how open-end review impacts data

Goals Of The Research:

1. Explore consistency between human reviewers
2. Analyze the effectiveness of an automated open-ended review platform
3. Assess the impact of data cleaning on the story

Hypothesis:

Open-end review is a key quality control method. As such we anticipate that removing respondents with poor open-ends will impact the data.

Survey Design

Methods

Through our annual Research on Research survey, The Harris Poll interviewed 11,469 US adults age 18+, across 13 different online opt-in sample provider blends from August 16, 2023, through August 31, 2023.

Due to the time intensity of open-end review, we selected a subset of the data for this experiment. Our research includes 2,696 US adults age 18+, across three of the sample providers.

The final datafile totaled 2696 IDs and included both qualified and previously deemed “fraudulent” IDs. Each reviewer (6 human, 1 automated) marked IDS for removal, which were then cleaned from their respective files. The result was 7 unique datasets, finalized based on the reviewers’ recommendations.

Data were RIM weighted in total (n=2696), and also separately by reviewer source, to population proportions from the Current Population Survey (CPS) 2022 for:

- Education
- Age by Gender
- Race/Ethnicity
- Region
- Household Income
- Household Size
- Marital Status

Individual weights were capped at 5 and 0.2.

The results presented are based on a gen pop survey of adults, 18+ in the U.S. with a 97% Incidence rate.

Findings may not be generalizable to studies with lower incidence rate or harder to reach audiences.

Goal 1: Explore consistency between reviewers

Human reviewers flagged about the same overall number of open-ends as invalid.

Volunteer 1

4%

Volunteer 2

3%

Volunteer 3

3%

Volunteer 4

5%

Volunteer 5

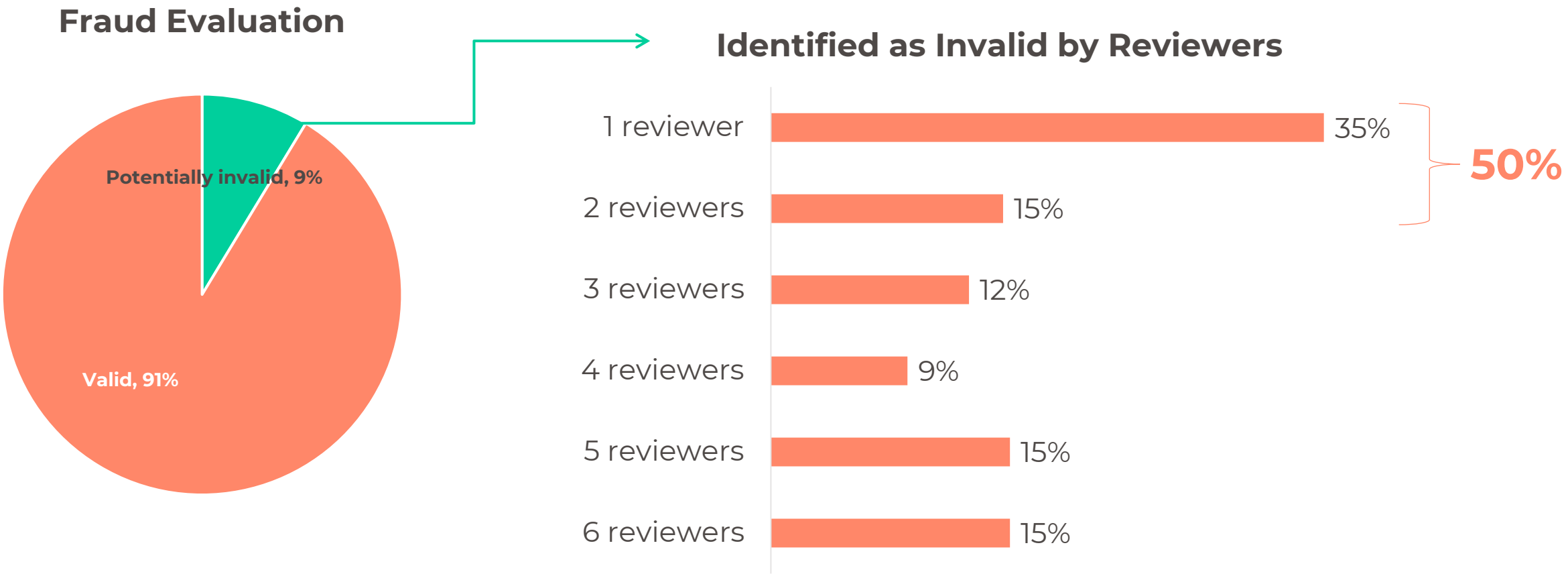
5%

Volunteer 6

5%

However, removing questionable respondents can be *highly* subjective.

Among open-ends identified as invalid, all reviewers agreed on just 15% of cases. And 1/2 of the potentially invalid responses were identified by just 2 reviewers.



What does a quality check with open-end removal look like?

Some Examples of Open-end responses

What is your favorite television show and why?	Issue	# of Reviewers
Yesterday and it was a great weekend and see you because of the best of luck and the other day I am	Does not answer question	6
My favourite hobby is a playing football with pele (x2)	Exact repeat	6
Good	Overly generic, common response	6
Great	Overly generic, common response	6
Very good	Overly generic, common response	6
Good vibes great company work I enjoy cool neag cool lit neat cool lit	Does not answer question	5
Who is that, lucky guy to have a lady like you crushing. maid coin doll check prize bullet metal reform bleak average suffer, meadow muffin dream expire wool want cloud wheel oxygen upgrade one rail come.	Does not answer question	5
Hfcvj	Random letters	5
Friends my lovely show	Odd grammar	4
television show is very unique quality and best survis	Does not answer question	3
Caso Cerrado, porque hablan espanol	Non-English response	2
My favorite television show is sure tank because it's	Typo, incomplete thought	2
i don't have personal preferences or feelings so i don't have a favorite television show however i can help you find information about popular tv shows if you'd like	AI Generated	1
Showmax because it is realiable and fun	Answers the question? Generic?	1
Amazon	Answers the question? Generic?	1

Goal 1: Key Findings

- **Its relatively easy to find consensus on valid responses (where all reviewers agree).**
- **And, Its relatively easy to find consensus on obviously invalid responses (where 5+ reviewers agree).**
- **But a significant portion of responses are questionable.**
  - Removing these questionable responses (where 1-4 reviewers agree) is *highly* subjective.
  - There is a risk that this subjectivity may inadvertently bias the data.
- **Adding a second reviewer may help reduce the subjective bias of open ends.**
  - Removing only open-ends where 2 people agree would mean less removals, but ideally would lead to more confidence that those open-ends truly are invalid.
- **Rules for removals (and non-removals) help improve consistency.** Examples might include:
  - Overly vague like good, great, very good
  - Random letters
  - Clearly does not answer question
  - None, n/a, DK are appropriate (even if disengaged)

Goal 2: Analyze the effectiveness of an open-end software review tool (OE Tool).

What is an open-end software review tool (OE Tool)?

- An OE Tool is an application that employs a score to indicate whether or not an open-ended response could be invalid. Based on the respondent's score they will either pass, fail, or receive an "undetermined" mark.
- Often times, a "fail" may result in an automatic kick out from the program if specified by the team or programmer.
- The OE Tool uses several factors to determine whether an ID should receive a flag. For example:
  - Bad language - words and expressions
  - Nonsense and garbage words - words not found in lexicon for the specific language
  - Robot submitted responses - text filled in automatically
  - Engagement - length of response compared to others
  - Repeated words percentage - percentage of 3 or more consecutively repeating words
  - Respondent duplication - identical responses

The OE Tool identified about the same number of invalid open-ends as a human reviewer.

Volunteer 1

4%

Volunteer 2

3%

Volunteer 3

3%

Volunteer 4

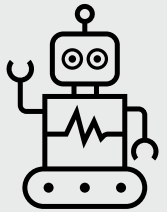
5%

Volunteer 5

5%

Volunteer 6

5%

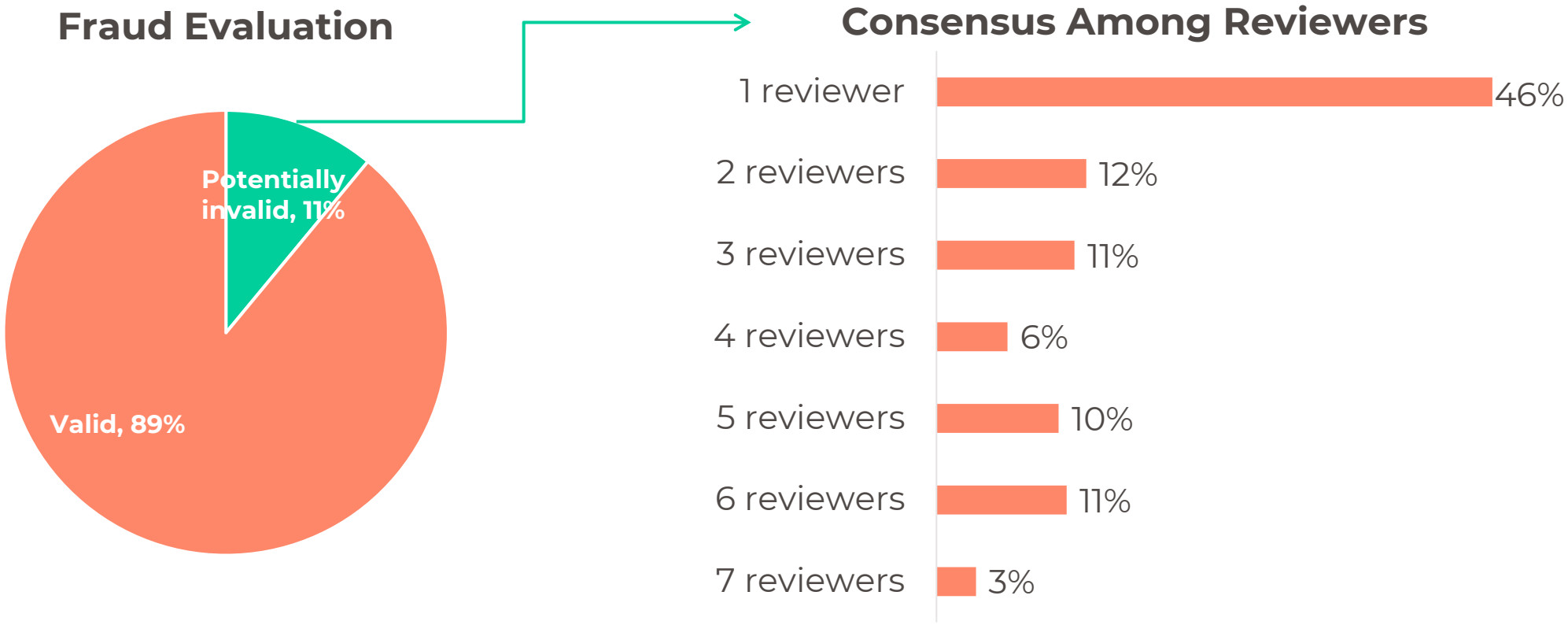


OE Tool

4%

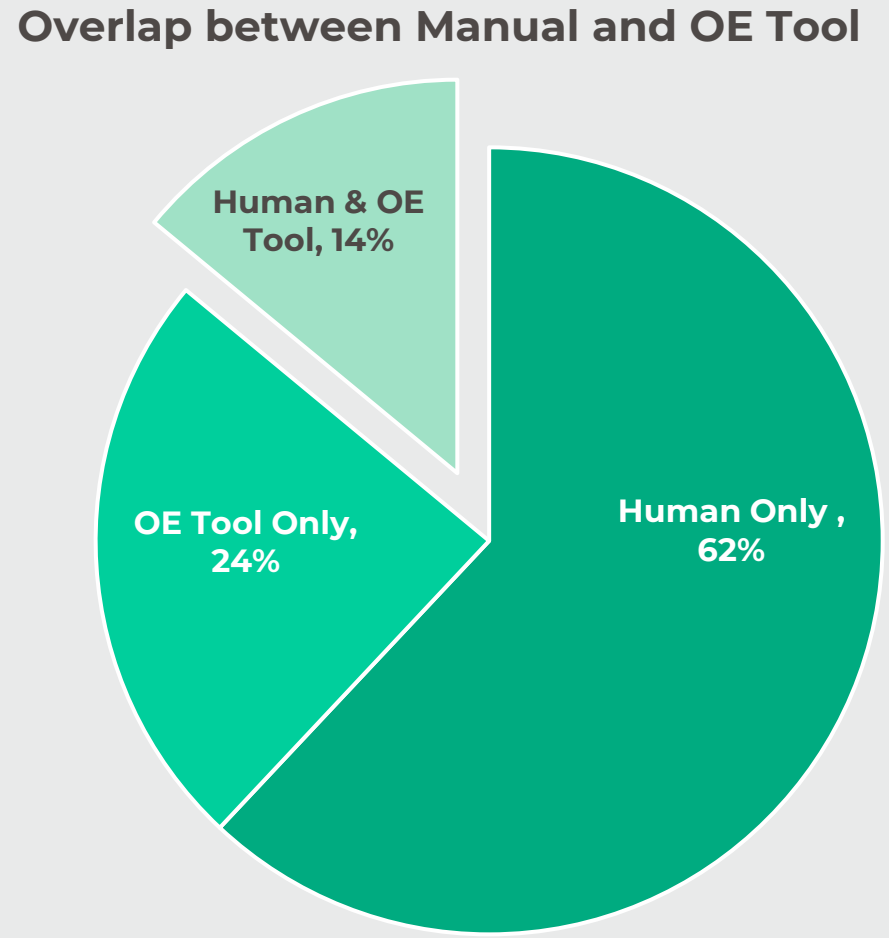
However, the OE Tool does not improve consistency in review.

The OE Tool increases the amount of potential invalid responses (11% vs. 9%), and almost half of those cases are identified as potential fraud by either only 1 individual or just the OE Tool.



Minimal overlap found

Out of 306 total IDs identified as potentially invalid by at least one reviewer or the OE Tool, only 14% were flagged as invalid by both sources.



Does the OE Tool get it right?

***Sometimes*....If it's repeated, too short (sometimes), or random/non-English words.**

What is your favorite TV show and Why?	Issue	# of Reviewers	OE Tool?
Yesterday and it was a great weekend and see you because of the best of luck and the other day I am	Does not answer question	6	No
My favourite hobby is a playing football with pele (x2)	Exact repeat	6	**Yes**
Good	Overly generic, common response	6	**Yes**
Great	Overly generic, common response	6	**Yes**
Very good	Overly generic, common response	6	**Yes**
Good vibes great company work I enjoy cool neag cool lit neat cool lit	Does not answer question	5	No
Who is that, lucky guy to have a lady like you crushing. maid coin doll check prize bullet metal reform bleak average suffer, meadow muffin dream expire wool want cloud wheel oxygen upgrade one rail come.	Does not answer question	5	No
Hfcvj	Random letters	5	**Yes**
Friends my lovely show	Odd grammar	4	No
television show is very unique quality and best survis	Does not answer question	3	No
Caso Cerrado, porque hablan espanol	Non-English response	2	**Yes**
My favorite television show is sure tank because it's	Typo, incomplete thought	2	No
i don't have personal preferences or feelings so i don't have a favorite television show however i can help you find information about popular tv shows if you'd like	AI Generated	1	No
Showmax because it is realiable and fun	Answers the question? Generic?	1	No
Amazon	Answers the question? Generic?	1	**Yes**
News	Disengaged	0	**Yes**
Star Wars	Disengaged	0	**Yes**

Goal 2: Key Findings

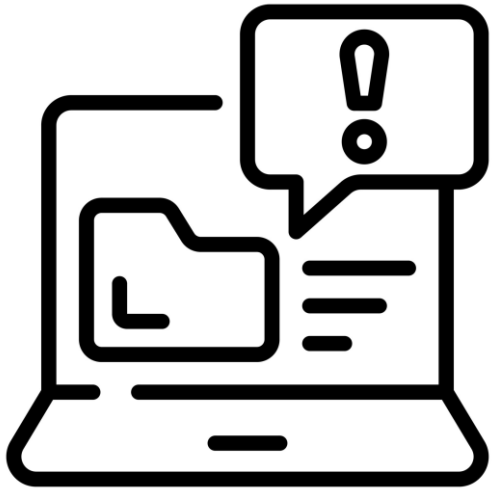
- **The OE Tool has *some* potential to act as a second reviewer.**
  - It can help to identify random letters, nonsense words, overly generic responses, and other potential invalid responses

- **The OE Tool did not improve consistency and could not replace a 2nd human reviewer**
  - It is too prone to flag legitimate responses that are too short
  - It is not able to identify if a response actually answers the question

Goal 3: Assess the impact of open-end removals on the story.

Open-end removals had no *practical* impact on the overall findings.

In this 15-minute survey w/ around 90 questions, just 2 had any significant impacts at the 95% confidence interval and 6 at the 90% confidence interval.



Questions with differences at 95% confidence Interval

2

Questions with differences at 90% confidence Interval

6

Open-end removals have the potential for a somewhat larger impact on subgroups.

In this research, open-end removals had a more substantial impact on males and respondents younger than 45.

MEN
Questions with differences at 90% confidence Interval
9

Under 45
Questions with differences at 90% confidence Interval
8

WOMEN
Questions with differences at 90% confidence Interval
1

45 and older
Questions with differences at 90% confidence Interval
0

While statistically significant, these differences are practically very small.

At the total sample level, removals led to a 1-2% shift.

Example 1: Total sample at 95% confidence								
What is your present religion, if any?	Sample (2696) (A)	Volunteer 1 (B)	Volunteer 2 (C)	Volunteer 3 (D)	Volunteer 4 (E)	Volunteer 5 (F)	Volunteer 6 (G)	OE Tool (H)
Muslim	4% BEDGF	3%	3%	3%	2%	2%	2%	4% EFG

At a subgroup level, removals led to as much as a 2-4% shift.

Example 2: Men at 90% confidence - Removals led to smaller percentage of gig workers								
Do you currently earn any money from a job that is considered gig work?	Sample (2696) (A)	Volunteer 1 (B)	Volunteer 2 (C)	Volunteer 3 (D)	Volunteer 4 (E)	Volunteer 5 (F)	Volunteer 6 (G)	OE Tool (H)
Yes	28% egf	25%	26%	25%	24%	24%	25%	28% e

Goal 3: Key Findings

- **Open-end removals had no practical impact on the overall findings for our study**
  - While there were some small differences, there were none that would be considered impactful for the purposes of our study

- **Open-end removals have the potential for a somewhat larger impact on subgroups**
  - Specifically younger respondents and males had more significant differences at the 90% confidence interval

- **Impact of open-end removals will be unique to every study**
  - Studies where 1-2% differences won't impact the story or where there is not much subgroup analysis may consider forgoing open-end review, or simply doing a very high-level review
  - Studies where 1-2% differences matter, will want to consider the impact of open-end removals
  - Studies with smaller base sizes, especially among key demos of interest, will want to consider the impact of open-end removals

Conclusion

Future Research, Limitations, & Caveats

This research was conducted based on a high IR, gen pop study. It may not be generalizable to studies with low IR or harder to reach audiences, where fraud is more apparent.

Ideally this research could be extended to different audiences, including those known to have higher levels of fraud, to determine the impact of open-end review and removals among those audiences.

We cannot say if the data removal is better or worse. Just that there are differences from the original to the cleaned file.

Ideally, we would add in an additional layer of verification to determine if respondents identified with invalid open-ends truly should have qualified and gauge their level of attention during the survey.

In addition to open-end review, we employ numerous quality checks and use trusted panel providers with their own layers of verification to ensure highest quality data for all our research.

Thank You

For more information, visit

theharrispoll.com

Andrea.Date@harrispoll.com
Julia.Nelson@harrispoll.com

/harris-poll /theharrispoll @harrispoll

The Harris Poll

APPENDIX

Bios

Andrea Date

Andrea Date stands at the forefront of market research and public service, embodying a unique blend of expertise that spans across the realms of policy, sustainability, and community engagement. As the Vice President of Research Methods and Services at The Harris Poll, Andrea has played a pivotal role in shaping the way organizations understand and interact with their audiences. Her work is characterized by a deep commitment to leveraging data for insightful decision-making, a principle that has guided her through a distinguished career in market research.

Andrea's academic foundation is robust, holding a Master's degree in Natural Resources Science and Management from the University of Minnesota-Twin Cities, complemented by a Bachelor of Arts from Carleton College. This educational background has equipped her with a nuanced understanding of sustainability and environmental policy, themes that have been recurrent in her professional journey.

In addition to her role at The Harris Poll, Andrea serves as a City Council Member for Woodbury, Minnesota, where she has been instrumental in driving initiatives that enhance community engagement and sustainable development. Her involvement in the city's Parks and Natural Resources Commission and the 2040 Comprehensive Plan Task Force exemplifies her dedication to fostering environments that prioritize both ecological and societal well-being.

Before her tenure at The Harris Poll, Andrea honed her skills as a Research Director at Material and Ravel, LLC. These roles underscored her ability to navigate complex data landscapes and extract meaningful insights, a skill set that she continues to leverage in her current positions.

Andrea's multifaceted career is a testament to her versatility and commitment to making a positive impact through research, policy, and community involvement. Her contributions to the market research industry and her community in Woodbury highlight her as a leader who not only understands the numbers but also the people and the planet they represent.

Julia Nelson

Currently working at Harris Insights and Analytics as a Senior Research Analyst, her role focuses closely on research for public release and thought leadership in the non-profit and healthcare space. At the Harris Poll she aims to promote and improve data quality, internal processes, and leverage her unique blend of research acumen, community service, and leadership.

Her research journey began at William Paterson University of New Jersey. At the university her work centered on an NIH-funded study which aimed to determine the efficacy of nutrition education and farmers' market voucher validation for folks accessing WIC services.. The work was later published under "Pilot Study of a Farm-to-Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) Intervention Promoting Vegetable Consumption."

Before joining formally entering the research space, Julia served as an AmeriCorps member at El Programa Hispano Católico, where she provide crucial support to local community members. Her efforts in trauma-informed care, barrier mitigation, and volunteer management developed a strong belief in the power of community and research to guide non-profit efforts.

Julia has a bachelor's degrees from The College of New Jersey, and a long history of volunteering with the local Interfaith Food Pantry and Girl Scouts of America.

Abstract

Is Human Review of Open-Ended Responses from Non-probability Online Sample Panels Needed?

One of the side-effects of incentivizing respondents to take surveys as part of an online panel is that inattentive respondents might be incentivized to provide minimal input, necessitating attention check questions for quality control (Curran, 2016). These inattentive respondents are more prone to things such as recall bias but also make up an important segment of the population that cannot be ignored (Alvarwz & Li, 2022). Along with inattentive respondents, there is increasing concern of fraudulent respondents, who intentionally misrepresent their qualifications, to complete and earn survey incentives. One of the most time-consuming and costly checks commonly used to assess inattentive and fraudulent respondents is a manual review of the respondent's open-ended questions. We explore the effectiveness of an automated open-end review platform in either replacing or supplementing human review of the open-end responses. Five reviewers will independently analyze open-end data from a nationally representative consumer omnibus survey in the United States and flag inattentive and fraudulent open-end responses. These responses will be compared to the automated platform responses in the same data set. The respondents flagged as fraudulent from each process will be compared for commonalities. Lastly, we will compare the cleaned datasets, with fraudulent respondents removed to the original dataset to determine the impact that open-end review and subsequent removals has on the results of the closed ended questions in the omnibus. We anticipate that commonalities in independent review will lead to a best practice guide in identifying fraud in open-ends and subsequent cleaning of data.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. Journal of Experimental Social Psychology, 66, 4-19.

Alvarez, R. M., & Li, Y. (2022). Survey Attention and Self-Reported Political Behavior. Public Opinion Quarterly, 86(4), 793-811.