

Double Dipping: A Review of Duplicate Responses Within and Between Online Panel Samples

Session: “Are You For Real? Bots, Fraud, and Careless Responses”

Matthew Deihl, Research Manager
Paul Johnson, Vice President

May 16, 2024



Hit Points

- Matthew Deihl, Harris Poll
- Duplicate Responses...
- What we are doing, and
- What we plan for in the future

Rough Script

Good morning, everyone – My name is Matthew Deihl. I’m a Researcher at the Harris Poll. And today I’ll be mulling over a specific type of quality issue in surveys – duplicate responses.

In this presentation, I’ll discuss how and why we see duplicates in our research.... what we are currently doing to combat against it. And what advancements we are considering for the future.

The Problem

Double Dipping: A Review of Duplicate Responses

- ❖ **Online Panels can be a great source for quick and affordable sampling.**
- ❖ **But Online Panels come with The Good, The Bad, and The Ugly.**
 - ❖ The Good – Engaged Respondents, who are vocal about their opinions.
 - ❖ The Bad – Fraudulent Respondents, who actively try to game the system for incentives.
 - ❖ The Ugly – Real Respondents, who may not be engaged with the survey, but represent an important part of the General Population.



Hit Points

- Rely on online nonprobability panels
- Fast & affordable.... Data quality issues
- Good, Bad, and Ugly
- Ugly are still real responses and represent GenPop

Rough Script

At Harris – we largely rely on online, non-probability panels for our sampling. Using these panels expedites our fielding periods and helps keep our budgets low. But these panels come with data quality issues in the responses. Respondents can be seen as ‘the Good, the Bad, and the Ugly’. *Good* respondents answer the survey in a timely manner, they are attentive to the survey items, and descriptive in their responses. *Bad* respondents are fraudulent. Either Bots or click farms, or just your average Joe who is just trying to pass your screener and get the incentive. The *Ugly* are real people and are largely being honest, though they may be less engaged and sometimes answer questions incorrectly. But they represent an important part of your audience.

The Problem

Double Dipping: A Review of Duplicate Responses

❖ Fraud Reviews:

- ❖ Illogical Choice Combinations
- ❖ Verbatim, Straight-line, and Speeding Checks
- ❖ Duplicate Reviews

❖ What is a duplicate response?

- ❖ Bots can accidentally run a script that duplicates its answers.
- ❖ Click-Farms can be copy/pasting from a list of approved verbatim responses.
- ❖ Many are real respondents who are simply members of multiple panels.



Hit Points

- Diff between Ugly and Bad
- Checks for Fraud – Illogical choice, straight-line, LOI - & Dupe Reviews
- Dupe review is not straight forward or efficient
- Dupes could be BOTs or Click Farms OR
- Real people who are members of multiple panels.

Rough Script

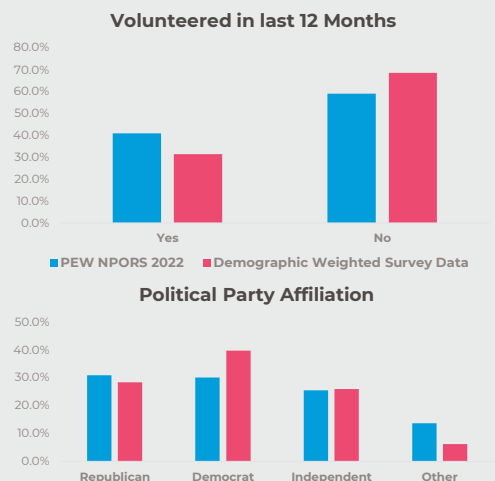
So – we need to figure out the best ways to distinguish the Bad responses from the Ugly responses and disqualify fraudulent ones . There’s many checks you can build into your survey to review for fraud – illogical choice combinations, straight-line checks, speeding checks. But reviewing for duplicates is not as straight forward. Duplicates can occur if a Bot runs the same script twice. Click-farms whose employees are following a script with approved verbatims. But most often times, it’s a real respondent who happens to click on a survey link twice, or is a member of more than one panel.

Harris' Annual Research on Research

Double Dipping: A Review of Duplicate Responses

- ❖ To assess the severity of this problem...
- ❖ The Harris Poll fields a 15-minute survey of over 11,000 US GenPop Adults, annually.
- ❖ This survey has multiple goals, such as
 - ❖ Updating benchmarks of US attitudes and behaviors on various topics, such as smoking habits, hobbies, and charitable donations.
 - ❖ Providing a safe place for methodological experiments.
 - ❖ Assessing sample quality from fourteen different Online Panel Samples.

The Harris Poll



Hit Points

- Frequency of Dupes?
- Annual “Research on Research” survey
- Update our benchmarks of attitudes and behaviors in the US
- Provides a safe place of methodological experiments
- Compare data quality between vendors.

Rough Script

Our team wanted to assess how frequent duplicates are occurring. And coincidentally – Harris holds an annual ‘research on research’ survey that aims to verify and update our benchmarks of US attitudes and behaviors on various topics. But also, it provides a safe space for us to implement methodological experiments and compare the quality of online panels. This round of the project used 13 online panels and collected over 11,000 GenPop responses for a 15-minute survey.

Methods of Identifying Duplicates

Double Dipping: A Review of Duplicate Responses

❖ Digital Fingerprinting (DF)

- ❖ Uses metadata from respondents' device, such as MAC and IP address, to create a unique ID for that device. Automatically terminates any additional responses from the same device.

❖ Manual Review of Data Similarities

- ❖ Reviewing certain key data points, such as same ZIP code and age, and seeing if other data entered have other patterns & similarities.

❖ Nearest Neighbor Analysis

- ❖ A new technique using Mahalanobis Distance of all ordinal, scale, and numeric data to aid Manual Reviews



The Harris Poll

5

Hit Points

- How to review for dupes?
- Digital Fingerprinting
- Manual review (Zip/Age/Verbatims)
- Developing a new technique – Nearest Neighbor Analysis.

Rough Script

There's a few ways we review our responses for duplicates. On all Harris surveys, we use a Digital Fingerprinting tool which encodes metadata from the device used to participate. If the survey is accessed more than once from the same device, the initial response is retained, but subsequent responses are automatically disqualified. However this is easily circumvented - a respondent just needs to use their computer for one response, and their cell phone another. We also manually review our qualified responses for certain similarities. We generally identify responses coming from the same zip code who are the same age, and then review for other data that might be 'suspiciously' similar. For this project, we also tried something new. My colleague and coauthor, Paul Johnson, wrote a script in R to calculate the Mahalanobis Distance of the survey's ordinal, scale, and numeric data points to identify 'nearest neighbors' who may likely be duplicates.

Digital Fingerprinting

Double Dipping: A Review of Duplicate Responses

❖ Pros:

- ❖ Automatic process behind the scenes.

❖ Cons:

- ❖ Can be fooled by VPNs or simply changing devices.

❖ Findings in Harris' Research on Research

- ❖ Out of 18,415 click-throughs, 11,469 responses ultimately completed a qualified response.
- ❖ Of the click-throughs, 15,882 unique devices were fingerprinted digitally.
- ❖ 608 of those devices (4%) attempted to take the survey more than once and had their subsequent attempts immediately disqualified.



The Harris Poll

6

Hit Points

- DF – Easy & behind the scenes
- Can be fooled – VPN or changing devices
- ~16k unique devices
- 4% of devices tried multiple times
- DF allows the first response to pass, but disqualifies all subsequent attempts

Rough Script

First – Digital Fingerprinting. It's a very easy process, running behind the scenes and doesn't add any extra effort to the research team. BUT, as I said before, it can be fooled. Using both qualified and disqualified data from Harris' Research on Research survey, we reviewed the digital fingerprinting metadata and found that nearly 16,000 unique devices attempted the survey, and 4% of these devices attempted to access the survey more than once.

Digital Fingerprinting

Double Dipping: A Review of Duplicate Responses

The majority of duplicates from multiple panels only came from just two panels – perhaps an honest respondent who is a member of two panels.

More rarely, respondents in 4, 5, and 6 panels are more questionable.

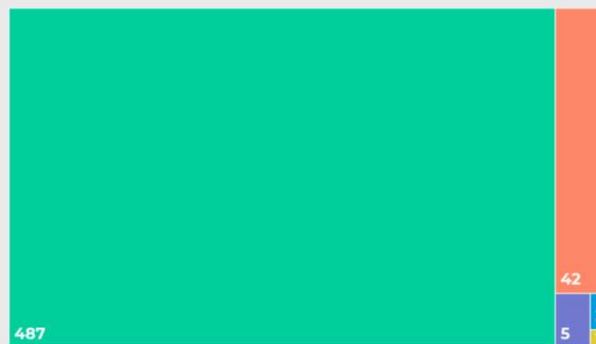
❖ Between-Panels Duplicates:

- ❖ 538 devices tried to access the survey more than once from more than one Panel.
- ❖ Every Sample Panel (13 total) used had at least one occurrence of a Between-Panel Dupe.
- ❖ Half of the panels had over 100 occurrences.
- ❖ Most Between-Panel Dupes only accessed the survey from two or three panels.

❖ **There is nothing preventing one person from being a member of two panels. And a panelist often does not know what survey is being offered when they receive an invitation.**

of Devices Accessing from Multiple Panels

■ 2 panels ■ 3 panels ■ 4 panels ■ 5 panels ■ 6 panels



The Harris Poll

7

Hit Points

- How many are coming from multiple panels?
- Of unique devices, 538 attempted from between panels.
- 90% came from just 2 panels – forgivable.
- Respondents who are on 2 panels don't have control for what survey is sent to them.
- All 13 panels had at least 1 instance of between panel duplicates.
- Half of duplicates only attempted from 2 or 3 panels.
- Only 8 devices attempted from over 3 panels.
- Those are most suspicious and generally caught by other fraud checks.

Rough Script

We wanted to see where these duplicate responses were coming from. We first traced devices coming from multiple panels. Of those 608 duplicate devices, 538 accessed the survey from more than one panel. 90% of these devices accessed the survey from two different panels. We assume that these respondents were members of two panels, and they simply do not have control over what surveys are sent to them. All of the 13 panels we used had at least one occurrence of a "Between-Panel Duplicate". And half of the panels had over 100 respondents who were members of more than one panel. On a good note - only 8 devices were found to be members of more than 3 panels.

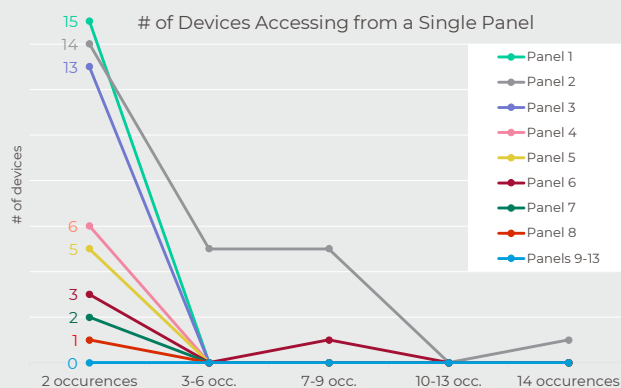
Digital Fingerprinting

Double Dipping: A Review of Duplicate Responses

Online Panels should be doing their due diligence to prevent duplicates from entering a survey. But not all panels have the best quality controls.

❖ Within-Panel Duplicates:

- ❖ Panels should be monitoring for duplicates accessing a link twice and preventing them from redirecting to the survey.
 - ❖ 8 of the 13 panels used had at least 1 occurrence of a single device accessing the survey two or more times.
 - ❖ Few panels had devices attempting more than twice. But one panel had 25 unique devices attempting access up to 14 times!
- ❖ **For these excessive duplicates, other fraud checks were able to identify the initial attempt and all attempts were removed from qualified responses.**



The Harris Poll

8

Hit Points

- Panels don't know if their members are participating in other panels...
- BUT they do know if their own members participate.
- 4 of our 13 panels did not have any duplicate attempts from their own panel,
- 8 panels had some instance of duplicates
- 1 panel in particular had devices attempt several times, up to 14 times!
- Some panels are better than others.

Rough Script

While panels have no way of knowing who is participating from other panels. They should be able to control for duplicates in their own panels. Reviewing digital fingerprints from the same panel, we found that 4 of the 13 panels had no occurrences of duplicate responses. AWESOME. But... most panels had devices attempting to access the survey twice. We can imagine reasons why this could happen legitimately. But some panels were obviously not monitoring for duplicate responses on their end. One panel in particular had a device attempting to access the survey 14 times. I should make one more positive note. Digital Fingerprinting only disqualifies the subsequent responses and retains the initial. But we found that these devices with excessive reattempts always had their initial response disqualified due to other programmed fraud checks.

Manual Reviews & Nearest Neighbors

Double Dipping: A Review of Duplicate Responses

❖ Manual review of Data Similarities

- ❖ Pros: Can catch the tech savvy fraudsters who can fool Digital Fingerprinting.
- ❖ Cons: Extremely effortful for limited return.

❖ Nearest Neighbor Analysis

- ❖ Pros: Automatic process after fielding
- ❖ Cons: Average computer takes multiple days to run the analysis.

❖ Using Mahalanobis Distance calculations across 48 questionnaire items - scale, single and multi select, and numeric entry - we identified response pairings with highly similar data.



The Harris Poll

9

Hit Points

- After DF, we manually review qualified responses
- This can be quick for surveys with low quotas, but
- the Research on Research study had over 10k!
- We filter on respondents from the same zip who are the same age, and compare their verbatims
- “What is your favorite TV show and why?”
- A response with same or similar answer is a dupe – Hannity & Tucker Carlson
- To streamline this task – we developed Nearest Neighbor reviews.
- Calculated the geometric Mahalanobis Distance across all 48 datapoints and paired responses that were most similar

Rough Script

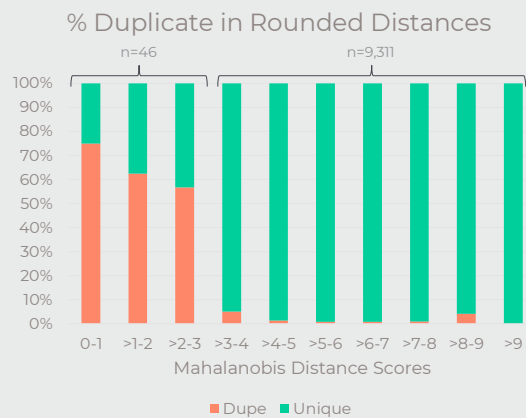
After digital fingerprinting comes the manual process of reviewing data. While this review may be quick for surveys with a few hundred responses, it becomes very time consuming for surveys that have 10s of thousands. We find the quickest method of the task is identifying respondents who are the same age and live in the same ZIP code, and we review their verbatim responses to see if they match. For our Research on Research project, we asked “What is your favorite television show and why?” Matching responses may be the same TV show, or similar TV shows with the same rationale. If one response said “Tucker Carlson because he shares my viewpoints” and another response from the same ZIP code with the same age said “Hannity because I agree with what he says”. We would likely consider this to be a duplicate. As you can imagine this review takes a lot of time. To streamline this task, we calculated the geometric Mahalanobis Distance across the survey’s 48 data points, we paired responses with a “nearest neighbor” response and scored their similarity.

Nearest Neighbors

Double Dipping: A Review of Duplicate Responses

❖ Findings

- ❖ Using Nearest Neighbor Analysis, 9,357 pairs of the qualified responses that passed digital fingerprinting were given a distance score ranging from:
 - ❖ 0 – identical data
 - ❖ 13 – least similar
- ❖ After manually reviewing, we found 134 of these paired responses had matching content in verbatims.
 - ❖ 104 were between two different panels.
 - ❖ 30 pairs were within the same panel.
- ❖ A major drop in dupes after a distance of 3.



The Harris Poll

10

Hit Points

- We identified over 9k potential pairs with Distance scores ranging from 0 to 13
- 0 is identical
- 13 is least similar
- Manually reviewed the pair's verbatims only
- Over half of pairs with distances of 3 or less were duplicates.
- Major drop off of similar verbatim responses for distance scores of over 3.
- Really – we only need to review 46 cases!
- Cut back on labor if we filter manual review down to distance scores of 3 or less.

Rough Script

After calculating these distances, we sorted the pairings from identical responses to least similar responses and reviewed the paired responses for duplicate verbatims. Out of 9000 potential pairs, we found that less than 50 had a distance of 3 or less. And for that distance range, over half of the pairs had similar or identical verbatim responses. A noticeable drop of duplicates were found in pairs with a distance of more than 3. For future projects, you can easily see how identifying nearest neighbors can aid the manual review and significantly reduce the level of effort.

Affects on Storytelling

Double Dipping: A Review of Duplicate Responses

❖ How does the data change with dupes?

- ❖ One of the primary concerns of fraud is its affects on the analysis we report on.
- ❖ To test this, we mimicked what the data would report by weighting duplicates by the number of occurrences of their digital fingerprint.
- ❖ Using a Z-Test, we tested at both 90% and 95% confidence to compare our qualified data against the data with mimicked duplicates.
- ❖ Not a single datapoint in the survey showed any significant differences between the two.



Hit Points

- How do duplicates affect our data?
- Created weights for our duplicate data to mimic what they would be if duplicates were allowed to qualify.
- Compared that data to our final qualified data
- Not a single data point showed a statistical difference between the two totals.

Rough Script

Out of curiosity – we also wondered how much these deduplication efforts really helped the data. We took the qualified data after deduplication and compared those totals to weighted totals that mimicked what the data would be if we had retained all of the duplicates. And we found that no data point in the entire survey showed any significant difference in the data's storytelling. So what does that mean...

Conclusions

Double Dipping: A Review of Duplicate Responses

❖ Tackling Fraud is an ongoing battle

- ❖ Fraudsters are finding new and better ways to navigate digital fingerprinting. Battling fraud is crucial to keeping it minimal.
- ❖ All DF duplicates were also identified as fraud through other in-survey quality metrics.

❖ For duplicate fraud...

- ❖ Our mimicked duplicate data found no changes to our storytelling. But duplicates should still be monitored for and disqualified.
- ❖ In the end, duplicates will invalidate the independence assumption on all of our statistical testing.

❖ Future Considerations

- ❖ Nearest Neighbor Reviews shows potential to help reduce the effort of manually monitoring for duplicates and keeping fielding efficient.
- ❖ This research used a GenPop audience, but is there more/less evidence of duplicates in Healthcare or Finance audiences?

❖ Is it worth the time and effort to review?

- ❖ Ultimately, it's for each researcher to decide.

Hit Points

- Tackling fraud is ongoing battle – but we are doing a good job.
- Duplicates may not have much affect on our findings, BUT
- They still invalidate independence assumptions in stat testing.
- Feel confident in your results, but still keep data clean.
- Considering building nearest neighbor calculations into our survey.
- Difference between audiences? Healthcare? B2B/Finance?

Rough Script

Tracking fraud is an ongoing battle. But we are doing a good job, keeping our responses as clean as possible. For duplicate fraud- sure, our deduped findings did not show any different to our duped findings, but in the end duplicates will invalidate independence assumptions on our stat testing. This just means that we can feel confident that if any duplicate responses slip between the cracks, it won't hurt our analyses. But we should still clean it up as much as possible. For the future – we at Harris are weighing the benefits of building a nearest neighbor process in our surveys to help reduce the time it takes to review data. And we are curious whether other audiences have different rates of duplicates. This research was on US GenPop – and these findings may be different for healthcare audiences or finance audiences. Is this something worth your time to review in your research? Ultimately, that's for you to decide...

Thank You

For more information, visit

theharrispoll.com

Matthew.Deihl@HarrisPoll.com



Abstract (pg 1 of 2)

Since the start of computer-assisted web interviewing (CAWI), researchers have been aware of the security concerns with duplicate data (Schmidt, 1997) occurring in their web-based data collection and the impact it has on their analyses. More recently, researchers have suggested that 20% of public use datasets have 5% or more of nearly identical observations (Kuriakose & Robbins, 2015). Whether it is due to intentional fraud, unaware professional survey takers, or bots entering scripted data, deduplication of survey responses is an important part of data quality review. The threat of duplicate responses grows as opinion research becomes more dependent on fast, inexpensive online sampling methods. For our research, we have reviewed nearly 18,500 responses to a US GenPop survey which used thirteen online panels with the aim of assessing how widespread duplicate responses are within and between the panels. We used two criteria to identify potential duplicates: the difference between observations as measured by a Mahalanobis distance, and the probability of those responses. We use both criteria as it is very likely to have identical responses when those responses are common, but much less likely to have identical responses when they are uncommon.

Matthew G. Deihl

Research Manager
Media Communications



Matthew G. Deihl is a seasoned research manager at The Harris Poll, specializing in large-scale survey operations and data validation. He holds a Master of Philosophy in Social and Developmental Psychology from the University of Cambridge and a Bachelor of Arts in Psychology and Philosophy from Notre Dame.

Matthew.Deihl@HarrisPoll.com
<https://www.linkedin.com/in/mattdeihl/>

Previously, Matt worked in education research – specifically Campus Climate and Career Pathways – enhancing survey methodologies and data collection strategies. He actively contributes to academic conferences and holds certifications in predictive analytics and SAS programming.

Edward Paul Johnson

Vice President
Advanced Analytics



Paul graduated from BYU with a M.S. in Statistics. He spent 15 years working in market research with a wide variety of experience including phone surveys, online surveys, and passively collected data. He loves finding innovative ways to combine behavioral and survey data with a specialty in discrete choice models.

Paul.Johnson@HarrisPoll.com
<https://www.linkedin.com/in/edward-paul-johnson-276a30244/>

He has volunteered and served on committees in AAPOR at both the local and regional level. His research efforts have been featured at AAPOR, ESOMAR, ARF, Insights Association, and Sawtooth conferences. He is most proud of his wife and two sons and loves playing games with them.

14

Abstract (pg 2 of 2)

We also compare observations using digital fingerprinting to identify respondents using the same device and compare how deduplication changes the bias in the sample compared to known benchmarks.

Those attending this presentation will be able to learn:

- A more sophisticated way to identify potential duplicate respondents
- Identify questions that work well in developing duplication criteria
- How duplicates potentially introduce bias compared to known benchmarks

Abstract Bibliography

Schmidt, W. C. (1997). World-Wide Web survey research: Benefits, potential problems, and solutions. *Behavior research methods, instruments, & computers*, 29(2), 274-279.

Kuriakose, N., & Robbins, M. (2016). Don't get duped: Fraud through duplication in public opinion surveys. *Statistical Journal of the IAOS*, 32(3), 283-291.