Validating Respondent Identity in Online Samples

AAPOR 68th Annual Conference
May 2013
Where We Started

Online research is huge, and growing
- Approx. $6B global spend in 2011 (ESOMAR 2012)
- Twice that of telephone based research

But...
- Online access panels continue to experience data quality concerns...
  - With sampling methods, and certain respondent behaviors
  - But also – the panelists themselves

The Problem: How do we know panelists are representing their identities truthfully online?
Validation - Confirming Respondent Identity

Everyone asks…

– How do we identify respondents who use false identities?
– How do they affect data quality and consistency?
– What can we do about this?

The Validation “Solution” – verify panelist identification through a consistent, structured process

But…

– How effective are these validation processes?
– How do they reduce the pool of “validated” survey takers?
– What impact do they have on the validity of survey results?
– How do “validated” vs. “non-validated” survey respondents differ?
First, What is Validation?

For our discussion, we’re talking about:
- Shorthand for “validation of identity”, using procedures to verify that people responding to an online survey are “real” people
- Applicable for panel-based respondents or intercept/River sample

Several companies provide services for online respondent validation, using the same basic process:
- **Collect** - personally identifiable information (PII)
- **Compare** – to national third-party databases (assumed to be comprehensive)
- **Categorize** – assign some evaluative code or label for validation status of their identity
Why Does This Issue Matter?

- Panel/sample providers ask …
  - Which one should I use?
  - How much does it cost?
  - Does it make a difference in resulting survey data?
  - What bias am I removing – and what bias am I creating?

- Buyers ask…
  - Why one process/supplier rather than others?
  - Does it make a difference in the validation? (Who am I missing?)
  - Does it make a difference in the data?
  - What bias am I buying?

- Everyone asks…
  - How do we know what validation does, or doesn’t, do?
The Hypotheses

- **Hypothesis 1**: You MUST validate
  - Validating identities and blocking unmatched volunteers improves sample and data quality.
  - *Test 1*: Look for variations in respondent data comparing those who can be validated versus those not validated

- **Hypothesis 2**: Just pick a solution, they’re all about equal
  - Validation methods are similar, and the choice of providers is less important as long as they have solid access to public records.
  - *Test 2*: Look for variations in respondent data comparing results by provider, analyzing match rates and data shifts

- **Hypothesis 3**: There may be a tradeoff in bias
  - Those who are unwilling to provide their PII for validation are different than those who will, and the process of validation itself may create data bias.
  - *Test 3*: Compare the data for those unwilling to provide personal info to those who are willing, both in terms of demos and survey content results.
The 2011 Research
To test these hypotheses, uSamp conducted 7,200 surveys during the first two weeks of January 2011

- The survey was a straightforward design – demographic, lifestyle, attitudinal and behavioral questions, with Census balanced sample
- We surveyed until we had 6,000 participants willing to provide their Name, Address, City, State, Zip code, and Birthdate (1,200 were unwilling)
- Data for all 6K with PII were submitted to four major validation firms:
  - Lexis Nexis
  - IDology
  - True Sample
  - Relevant Verity
The 2011 Validation Results

- Some intriguing stuff:
  - 17% of respondents refused to provide PII (name, address, DOB)
  - Average validation rate across the four suppliers was 83%
  - Validation rates were much lower for Under-25 respondents (*half* of other groups, esp. for males) and those indicating Asian or Hispanic ethnicity (20-30 points lower)
  - Respondents who didn’t validate exhibited different response patterns (i.e., tech use, home ownership, buying behavior) and data quality (50% more likely for straightlining and speeding)

Overall, *good news* and *bad news*:

- validation seemed to find “bad” respondents, BUT…
- those not validating were from the hardest to reach demographic groups, and also showed differences in consuming behavior – a troubling potential bias
So, on to 2012!

We decided to dig deeper a year later. Goals of the new research were:

- To see if validation processes had changed with providers’ claimed improvements since 2011
- To test the willingness of online respondents to provide PII as validation practices became more common
- To see if the impact on available sample composition was as significant as in 2011’s results
- To extend the analysis and comparison of validated vs. non-validated groups, especially as part of the overall sample achieved
- To gather new data about why some respondents refuse to provide PII, and what that might mean for general online research looking forward
2012 Survey Conduct and Results
The 2012 Survey – Planning and Implementation

We conducted essentially a duplicate project, with minor changes:

- Respondents asked at the outset to provide PII for name, postal address, and date of birth; email address asked at the end (since some services use this for validation); allowed to continue if refused
- Same fielding period, sample composition, general distribution among Census demographics for completed surveys, and survey structure/content
- Added questions about respondents' views of online privacy, Internet use, and comfort in purchasing online
- Collected 7,435 completed surveys, with mean completion time of just more than 5 minutes
- PII sent to the previous verification services, with one new provider added.
- All have somewhat different methods for the validation process – future research we plan to conduct will address a comparative evaluation.
The 2012 Survey – PII Provision

Roughly one in four respondents refused to provide full/any PII.

- Higher rate than the 17% seen in 2011.
- Adjusted for the inclusion of email address requested (not used in 2011), the rate was still more than 19%.

Although many statistically significant differences were shown between those supplying PII and those not, in many instances those differences were quite small (i.e., less relevant behaviors such as reading a Sunday newspaper). The more important findings:

- Those under age 25 and over 60 were less likely to provide PII
- Respondents identifying as Asian and Other refused at a higher rate
- Approx. 70% of those “very worried” about online privacy supplied PII, vs. 85% generally
- Similarly, less frequent online purchasers showed similar lower PII rates
The 2012 Survey – Validation Outcomes

All respondents who provided PII were sent to the validation services.

- Different categorization returns, from yes/no, to full/partial and 1-5 scale.
- Across the service providers, the percent of the sample clearing the highest bar for validation varied from 65% to 95%, with an average of 85%.
- Essentially the same as in 2011.

As with PII provision, validation rates showed similar patterns.

- Approx. 87% of 18-25 validated, 98% of 35-44, and 99% of 45-59
- Just 90% of Asians validated, vs. 96% of Hispanics and 97% for Caucasian and African Americans

No differences seen in some interesting dimensions for home ownership, having a bank account or credit card, or the three privacy questions.
The 2012 Survey – Data Quality

One new finding, very relevant to data quality.

- Respondents who failed verification were *twice* as likely to fail at least one quality check in the survey.
- Typically inconsistency between agree/disagree attitudinal battery questions sequenced in the survey, such as importance of brand names for purchase intent.
- 18% of non-validated vs. 4% of validated failed such checks.
The 2012 Survey – Data Quality cont’d

On the issue of excluding up to 25% of the end sample as non-validated respondents, and possible effects on end sample demographic makeup, as well as for answers to behavioral and attitudinal questions:

- No changes of any appreciable magnitude seen – some demographic distributions affected, but to a minor extent. Validation applied to a large sample didn’t skew the overall demographic balance.

- No meaningful differences in technology product ownership or for most of the attitudinal measures.

- Removing approx. 25% of the sample due to the validation process had no impact on the percent of respondents failing one or more response quality checks.
So What Did the 2012 Project Really Show?

We undertook to augment the 2011 study, and understand the dynamics of respondent validation in more detail. To that end:

PII:

- The higher PII refusal rate in 2012 was surprising. After looking into several possible influences, nothing clear emerged as to why this was.
- We did notice that those with elevated privacy/sharing online concerns refused PII much more frequently.

Validation:

- Validation rates within each of the services used increased consistently between the two studies.
  - Gain for each ranged from 5 to 11 points.
  - All substantially increased validation for 18-24 year-olds, Hispanics and Asians
- Possible explanations:
  - Validation services have improved their methods
  - Widespread use of validation is discouraging fraudulent online survey activity
So What Did the 2012 Project Really Show?

Finally, and most importantly: what was the impact of the validation process on sample composition and the potential for bias?

The combination of refusal to provide PII and the validation process itself resulted in about one in four prospective respondents being removed from the sample.

- Removed respondents were “different” in a number of ways – younger, more likely to be Asian, more concerned about privacy, generally spending less time online.

- As it turned out, these differences, while significant, were not of sufficient magnitude to seriously alter the targeted demographic distribution of the sample.

- Nor were there major changes in the behavioral or attitudinal measures in our short survey.

- Despite losing people from the “hardest” groups to get online, the overall character of the sample was little changed.
Conclusions

Validation of respondents is one of several responses to perceived data quality questions about results from online sample.

The findings of this study and the previous 2011 show that validation is something of a mixed bag.

- The process roots out some supposed “bad actors”.
- But these same people are easily caught by other quality checks and even if not, it’s not clear that they exist in numbers great enough to actually influence a survey’s results.

On the other hand, validation processes seem to be screening out a significant number of willing respondents only because they are reticent about sharing their personal information online.
Conclusions, cont’d.

In the end:

Fair to ask whether the expected benefits of validation are sufficiently great to balance the loss of as much as a quarter of the sample for a study – and a higher percentage among certain demographic groups?
Contact Information

Scott Worthge
Vice President
uSamp

scottw@usamp.com