

Respondent Validation: So many choices!

Does choice of vendor make a difference in respondent validation and how does it impact the overall sampling frame and data quality?

Melanie Courtright and Chuck Miller

DMS Insights, a uSamp Company
19111 North Dallas Parkway, Suite 350, Dallas, TX 75287

1.0 Background

In the digital world, the quality of online research data has been questioned time and again. Therefore, it is of significant importance for researchers to validate survey respondents in order to reduce errors and ensure data quality. As a result, a variety of methods and techniques have been instituted by companies who claim their process is better than others. However, the variation in the validation process used by different providers along with the cost associated with validation and other related complexities, makes it a daunting task for researchers to evaluate and determine if there is a real need for validation and the resulting difference in the survey data.

For companies that do validate respondents, because the validation effort is essentially meant to improve data quality, it becomes critical for the buyers to make sure that any potential bias is not being introduced by eliminating key demographics or hard to reach populations.

In this paper, we will look at sampling frames and data of valid vs. invalid respondents to determine the extent of variation along with evaluating data by providers to compare differences in the match rate and data shifts. Additionally, respondents who are unwilling to provide their identification information will be profiled and compared against respondents who shared their identification information to learn differences and find out if the validation process itself may produce data bias.

2.0 Methods and Design

An online survey that focused on demographic, lifestyle, attitudinal and behavioral questions was administered with a total of 7,200 respondents during the first two weeks of Jan 2011. Out of 7,200 respondents, 6,000 respondents provided their identification information which comprised of Name, Address, City, State, Zip code, and DOB (1,200 respondents were unwilling to share this information).

2.1 Survey Execution

The recruitment of respondents was controlled for age, gender, income and ethnicity to ensure even distribution. About 60% of the sample comprised of the company's own panel members and the other 40% was contributed by third party sources.

2.2 Validation Process

The Personally Identifiable Information (PII) gathered from the 6,000 respondents who were willing to provide it was submitted to four major validation firms that included Lexis Nexis, IDology, True Sample and Relevant Verity. The process and information received back varied by provider (data is masked and randomized throughout):

Provider A: Analyzed information in greater detail and reported whether or not they found a match for each of first name, last name, address, date of birth and assigned an overall validation score of 1- 5. The below validation score was used for categorizing the respondents in three major groups listed as follows:
5 = valid, 4 = partial valid, < 4 = non-valid

Provider B: Matched respondents on correct first name, last name, address, city, state, Zip, DOB and assigned a decision flag of success, failure, partial for each case. Additionally, codes with additional information for records that matched but had exceptions were also indicated. The decision flag was used for analysis purposes.

Provider C: As compared to provider A and B, provider C returned less information and left blank fields if address or DOB was not verified. For analysis purposes, non-valid were respondents whose address or DOB did not match. Partial valids were respondents whose DOB matched but address did not match and valids were respondents whose address matched.

Provider D: Sent least amount of information and had categorized respondents as either valid or invalid.

3. Findings

Responses to the attitudinal, behavioral, lifestyle and demographic questions along with identification and validation information were analyzed to seek the below insights.

3.1 Validation overview by provider, panel source, and ethnicity

Provider: The percentage of validated respondents varied across all the providers. Provider A and D had highest number of validated respondents at 86.2% and 87.4%. In comparison, provider C had least validated respondents (78.4%).

Historically, since the younger demographics are hard to reach, recruit and retain in the panel, this demographic criteria was compared across all the four providers to determine the proportion of validated respondents in the 18-24 age group. Nearly 50% of respondents in the 18-24 age group were validated for provider A (49.2%) and D (53.4%).

Panel Source: Respondents from the third party source did not subscribe to a research panel and were analyzed separately to review the variation in the validation results. Similar trends were observed where provider A (74.6%) and C (77.5%) had most validated respondents for the entire sample in the 18-24 age group.

Ethnicity: After comparing the total sample across different providers, it was clear that Hispanic, Asian and the other ethnic groups included the least number of validated respondents in contrast to White, African American and Native American ethnic groups (see Table 3).

3.2 Overall sample comparison for technology, attitudes and behavior

To further understand the differences in sample composition, the overall survey sample, respondents that provided personal identification information, respondents who refused to provide personal identification information along with invalids were reviewed for key demographics such as gender, age, income and ethnicity.

While there were no major gender ratio differences for validated respondents, Males comprised 64% of the invalid sample. The majority (78.5%) of the invalid population belonged in the 18-34 age groups and were predominantly non-whites. No prominent trends were seen between different samples for income.

When the demographics of sample that refused to share personal identification information was studied, it was found that higher percentage of females (53.9%) and respondents in the 18-24 were unwilling to share their information.

Technology: Additionally, there were significant differences regarding ownership of technology, attitudes and behaviors for valid, invalid and refused respondents. In totality, invalid respondents have a higher rate of owning iPhone (42.6%) and smart phones (58.1%) whereas refusers do not own much technology.

Attitudes and Behaviors: While valid respondents are less cautious about sharing information online, refusers overall seem be cautions regarding sharing information online and are conservative about making purchasing decisions quickly. Additionally, they don't care about brands and are worried about the environment.

On the other hand, Invalid respondents make quick purchases and are not price conscious. 72.1% of invalid respondents hold passport and 52.2% of them own their own home (see Table 4).

3.3 Data Quality

With the purpose of gauging the data quality of valid vs. invalid respondents, the data was analyzed to identify the percentage of straight liners, speeders and the ones that get caught in data traps. It was observed that the quality issues were obvious (31.8%) in non validated respondents as compared to (19.8%) valids, indicating that non validated respondents include more cheaters (see Table 5).

4. Discussion

- Dig deeper: While it is imperative to begin categorizing respondent into three major categories, it is evident that the invalids comprise of two sub categories that include the young, non-white people who are likely to be foreigners and might legitimately not have public records in spite of of being good survey respondents.
- What concerns them: Refused respondents include both good and bad survey takers. It is unknown if they refused because of privacy concerns or any other reasons.
- Dealing with Scammers: The major concern is regarding respondents who intentionally provide false information to be eligible to register, participate or be able to take the survey more than once. This group is likely to include scammers who may be getting the information from phone books and eventually get invalidated while providing the relevant DOB.

- It is all about money: While validation helps to reduce error, it costs to validate each person irrespective of the final status of being a valid or non-valid. Therefore, most companies have to over sample and include additional members to maintain the panel size while accounting for non valid respondents.
- Getting hold of them: The validation process gets particularly complicated with the younger age group (18 -24) as they are not only hard to find and recruit, but are equally difficult to validate.
- Impact on traffic: It is likely that the validation process may invalidate some good respondents who could provide valuable insights. In addition to impacting the overall data, it wastes traffic which has cost implications.
- Whom to choose: Every provider uses a unique validation methodology. Therefore, there is no common platform to compare and evaluate. However, it is crucial that the provider has solid access to public records to make sure that we do not lose any valid respondents.
- Bottom line: Validation helps, especially to eliminate bad scammers who register for pure monetary reasons and muddle the quality of survey data.

5. Further Research Considerations

- The online research industry along with the providers should focus on creating a standard validation methodology to ensure that respondents are evaluated consistently. In the absence of consistency, companies who validate should thoroughly test and compare providers before choosing a long-term partner.

For further information on this and other online sampling expertise, please contact:

Chuck Miller | President, DMS and Chief Research Officer, uSamp | chuck@dmsinsights.com

Melanie Courtright | Senior Vice President, Client Service | melanie@dmsinsights.com

Table 1. Respondent validation by provider

TOTAL SAMPLE	A	B	C	D
% Validated	86.2%	80.8%	78.4%	87.4%
% Partially Validated	2.3%	2.1%	13.2%	n/a
% Address match	88.5%	86.8%	86.8%	n/a

Table 2. Respondent validation by provider and age group

18-24 YEAR OLDS	A	B	C	D
% Validated	49.2%	36.6%	33.4%	53.4%
% Partially Validated	8.8%	5.9%	32.0%	n/a
% Address match	59.0%	46.7%	55.1%	n/a

Table 3. Respondent validation by provider and ethnicity

TOTAL SAMPLE	A	B	C	D
% White Validated	79.9%	74.6%	72.8%	82.2%
% African American Validated	74.1%	69.8%	65.6%	76.9%
% Native American Validated	74.3%	69.7%	71.3%	76.2%
% Asian Validated	56.2%	45.7%	42.0%	59.4%
% Hispanic Validated	63.6%	53.0%	52.8%	65.8%
% Other Validated	65.2%	56.1%	60.9%	68.1%

Table 4. Overall sample comparison

	All Sample	Sample with PI	Sample with PI (Valid) only)	Valid A	Valid B	Valid C	Valid D	Invalids	Refused
Male	47.5%	47.9%	47.1%	46.6%	46.8%	47.0%	46.8%	64.0%	46.1%
Female	52.5%	52.1%	52.9%	53.4%	53.2%	53.0%	53.2%	36.0%	53.9%
Ages 18-34	28.0%	25.6%	17.1%	18.0%	19.0%	19.8%	20.4%	78.5%	37.5%
Ages 35-54	39.2%	40.5%	43.0%	43.0%	42.8%	42.4%	42.4%	18.8%	34.8%
Ages 55+	32.6%	34.0%	39.8%	39.0%	38.1%	37.8%	37.3%	2.6%	27.8%
Income Under \$50K	38.0%	37.2%	35.7%	36.3%	36.8%	36.3%	36.7%	32.4%	41.2%
Income \$50-\$100K	40.7%	42.0%	43.3%	42.9%	42.6%	42.5%	42.6%	41.5%	36.7%
Income Over \$100K	21.2%	20.8%	21.2%	20.7%	20.5%	21.1%	20.8%	26.0%	22.2%
White	72.5%	74.8%	80.3%	79.2%	79.0%	78.1%	77.8%	47.9%	64.1%
Hispanic	8.6%	8.3%	6.2%	6.7%	6.6%	7.1%	7.1%	19.5%	9.6%
AA	13.3%	12.6%	10.8%	11.2%	11.5%	11.4%	11.5%	22.1%	15.6%
Other	1.9%	1.8%	1.4%	1.5%	1.4%	1.6%	1.6%	4.3%	2.2%

Table 5. Sample comparison for Attitudes and Behavior

TOTAL SAMPLE	Valid A	Valid B	Valid C	Valid D	Invalid	Refused
Shop around before buying	5.36	5.28	5.31	5.29	4.78	4.78
Impulse buyer	3.65	3.65	3.71	3.66	3.69	3.28
Price is more important than brand	4.93	4.91	4.95	4.92	4.33	4.39
Brand is more important than price	3.60	3.68	3.70	3.70	3.86	3.22
Not worried about environment	3.38	3.42	3.43	3.37	3.29	2.98
Concerned about global warming	4.19	4.19	4.34	4.38	4.04	4.09
Cautious about sharing online info	4.61	4.61	4.82	4.84	4.69	4.93
Hold passport	55.4%	53.0%	53.4%	53.9%	72.1%	58.3%
Own your own home	35.1%	37.0%	42.5%	41.6%	52.2%	32.0%
Left-handed	13.1%	14.9%	12.6%	12.4%	8.1%	8.6%

Table 6. Data quality comparison

TOTAL SAMPLE	A	B	C	D
% Valid with quality issue	20.4%	19.4%	19.1%	20.2%
% Invalid with quality issue	31.3%	30.7%	32.8%	32.3%