



ADOPTING A MULTI-TIER APPROACH TO DATA QUALITY IN ONLINE SURVEYS

sago.com



Data quality has been and will continue to be of paramount importance in marketing research and, in particular, in online surveys – and for a good reason: **Without reliable data, the analysis, on which critical business decisions are made, will not be accurate.**

Yet, there are no official industry standards on how to systematically approach identifying and/or removing “suboptimal” respondents from a data set. This is probably because, for each project and for each researcher, factors and risks vary considerably depending on research objectives, questionnaire design, methodologies used, panels utilized, and respondents targeted. So, identifying and evaluating “sub-optimal” responses becomes as much an art as it is a science.



That said, here at **Sago** we are pragmatic people, and below are some guiding principles and practical approaches that we apply and recommend in consultation with our clients.

FIRST THINGS FIRST!

DATA QUALITY IS A SHARED RESPONSIBILITY

Yes, we are in this together! Typically, the researcher drafts the questionnaire. It is evident that the quality of the survey questionnaire, the primary instrument to collect survey data, directly correlates with quality of data. In well-designed surveys, sub-optimal behavior occurs at low rates. So, reducing survey length (to no longer than 25 minutes), reducing redundancy and difficulty level, and deploying techniques to minimize certain behaviors will all help reduce sub-optimal respondent behavior. Sample providers or panel companies are responsible for maintaining a high-quality panel and for procuring a good quality sample for surveys. Survey programmers are responsible for consultation on how to optimize quality of programmed surveys.

Lastly, we expect survey respondents to be attentive to the survey and provide honest responses. All parties need to deliver on their shared responsibilities to achieve the highest quality data set.



'MULTI-TIER' APPROACH TO IDENTIFY SUB-OPTIMAL DATA



Removing responses based on a standalone criterion may result in over-elimination and may impact the sample representation negatively. Sago employs a “multi-tier” approach to identify sub-optimal data. Such an approach takes a little more work but it ensures we are balancing suitability of responses with human behavior. Before removing a qualified respondent from your data we examine a number of factors, such as the severity of sub-optimal behavior, survey length, design, complexity altogether with the impact on data representation.

Below We Discuss Some Commonly Used Criteria, How We Measure Them, and How We Might Evaluate Them:



Duplicate Respondents

Duplication may occur for multiple reasons, for example, a single respondent can be a member of multiple panels. In another scenario, especially in emerging markets, shared computers, internet cafes and/or kiosks may be more prevalent. In addition, when multiple respondents connect from the same company, campus, doctor's office, or household, a shared IP address may give the appearance of duplicate respondents. Sago uses an industry leading digital fingerprinting software that looks beyond IP address to

identify duplicate computers and prevent survey respondents from taking an online survey more than once.



Geo Location

Our digital fingerprinting software also identifies the geographic location of the user by country, state/region, city, U.S postal code, metro code, latitude and longitude information, and verifies that the IP address is originating from where the individual is located. If not, the respondent is flagged for further review.

It is worth noting that sometimes, due to legitimate reasons, such as travel, respondents take surveys from a different country to where they may 'belong' according to our panel. An individual respondent who is occasionally in another country may be considered safe; hence, proper discretion is used.



Speeders

A speeder is someone who completes a survey in what is perceived to be an unrealistically short period of time. We must consider that some surveys with complex routing logic may exhibit a wide range of expected survey completion times. Sago recommends 40 percent of median survey length as a threshold. Anyone completing the survey in 40 percent or less of the overall median survey length is flagged as a speeder.

Speeding behavior is evaluated individually and at a high level. Certain demographic groups, such as younger participants or participants who have an education level of college graduate or above, often exhibit speeding behaviors. By automatically excluding such speeders, in some instances, we risk excluding key demographic groups and potentially biasing the data.



Straight-Liners

Within table-based or attribute-based questions (grids), there exists the possibility that a respondent may simply click answers in a straight line in their haste to complete the survey. Sago employs the following method to identify straight-liners: If a respondent straight-lines 3 grids (all responses in each grid) for a survey longer than 25 minutes (or 2 grids for a survey shorter than 25 minutes) in median length then s/he will be flagged as a straight-liner. These straight-lining rules are not applied to attitudinal questions, where straight-lining of responses may be legitimate.

There are evidences that the length of the survey, number of attributebased rating questions (grids), and scale of the grids have an impact on respondents' straight-lining behavior. In order to mitigate this issue, we recommend keeping the

survey to 25 minutes in length or shorter and the number of grid type questions to a minimum. Creative representation of grids, such as card sorts or slider scales, can be viable alternates and can help minimize straight-lining.



Open-Ended Responses

A respondent who rushes through a survey is apt to take short cuts and may type in gibberish or nonsense when prompted for an open-ended response. This may also indicate that the survey is being answered by a bot rather than a human being. The best way to catch this behavior is to review the open-end responses. To the best extent possible, all open-ended responses are reviewed and any “nonsense” or “inappropriate” response is noted as a “soft” strike. If a respondent accumulates multiple strikes of this particular type, or uses egregious or

offensive language, that case (i.e. the entire respondent record) is removed.

Depending on the length, design, and the number of open ended questions, respondent behavior may vary from respondent to respondent and from survey to survey. In certain scenarios, a respondent may not have a specific comment for those questions, meaning that comments such as “n/a,” “don’t know,” or a blank open-end may be valid.



Red Herring Questions

Within a list of possible answer choices, a nonexistent/ impossible answer choice may be programmed as a ‘red herring’ question. The benefit to implementing such questions is that they are helpful in quickly identifying spurious respondents. The downside of using this technique is the risk of irritating legitimate respondents. (Please note:

Sago does not add red herring questions to client surveys, unless specifically requested by our clients).

The above are just some of the multiple methods and techniques used to identify data that may impact quality. In general, any respondent exhibiting unusual behaviors or providing inconsistent or inappropriate responses are marked as suspicious and, in some cases, removed from the data set after a thorough review.



Controls Don't Have to be Manual

Our programming team often builds 'real-time' data-cleaning measures during the programming phase. This can aid the lead researcher by identifying 'flags' based on common data cleaning criteria, such as speeding, straight-lining, and red-herring questions; and with a slightly more custom script can also identify inappropriate verbatim.



Summary

Since each survey design, complexity, length, and target is unique, the person cleaning the data needs be extremely thoughtful about how these “rules” are applied. Rather than making a decision based on a single measure, a holistic approach to analyzing the survey responses, along with the added application of experience, knowledge of the topic, and the target group, creates the right balance.

A Word About Sago

Sago is the global research and data partner that connects human answers to business questions. Combining our history, global reach, and innovative spirit, we enable our

clients to solve business problems through extensive audience access and an adaptive range of qualitative and quantitative solutions. We help our clients understand what their customers want and demand – empowering them to make decisions with confidence. As a partner to our clients, their clients, and the industry, Sago seamlessly connects businesses to key insights.





Adaptive solutions, confident decisions.

Let's Work Together

connect@sago.com

sago.com