

Using Analytics to Improve Data Quality in Survey Collected Data

Survey data is different in nature from other types of empirical data in that its fundamental basis is subjective and based upon the particular frame of reference and current state of mind of the respondent completing the survey. Further the basis of survey information is what a particular respondent intends or perceives as opposed to how that have actually acted. As such, one must consider analytics in a different light to enhance the quality of survey data.

The ability to use analytics to improve the quality of survey data begins with establishing a consistent data collection methodology. A high degree of consistency is required in order to identify and correct erroneous data in the survey.

DATA COLLECTION METHODOLOGY

- **Anonymous Data** – Anonymous Data is that data collected without reference to an individual. There may be geo demographic components of the survey that relates the response to a group for measurement and reporting of survey results. In general, analytics to enhance data quality cannot be effective when applied to anonymous data because no specific trend information exists for the respondent.
- **Authenticated Data** – Authenticated Data is that data collected with reference to a specific individual tracked over multiple surveys. Geo Demographic data is almost always associated with Authenticated Data to allow enhanced understanding of the individual, classification of the individual into a larger group(s) and to establish a comparative standard between the individual and the groups(s) to which they belong.
- **Base Line Questions** - Base Line Questions are those questions that remain constant between surveys. The purpose of Base Line Questions is to measure the consistency of the responses over time. Significant changes in the response to the Base Line Questions may indicate a material change in the views, sentiment, environment or attitude of the respondents.
- **Amplitude / Magnitude Questions** – Amplitude Questions are those questions that measure a respondent's perception of what constitutes large medium and small. For example, an individual in Accounting/Finance would have a very different perspective on what constitutes a large monetary amount than an individual in Client Services because of the numbers they deal with on a daily basis.
- **Exploration Questions** – Exploration Questions are those questions designed to discover the views or perception of the respondent. The nature of the Exploration Questions varies from survey to survey and is dependant on the type of information that the surveyor is seeking to gather. As with the Base Line Questions, it is important to, define the responses quantitatively to feed the analytics process. Qualitative questions may be required in this section to gain a full understanding of the survey issue(s), however the review of this data is manually intensive and not readily adaptable to analytic scrutiny.
- **Frequency** – Frequency of surveys depends upon the length of the survey and the perceived volatility of the survey climate.

DATA ORGANIZATION

- **Individual** – A unique identifier links responses from an individual respondent over multiple surveys.
- **Group (Segment)** – Individuals are grouped together based on common characteristics to better understand the dynamics of the group. Organizational structure, as in the case of a company, or geo demographic characteristics are generally the criteria to group the individuals. Note that it might be appropriate to classify an individual into more than one group.
- **Structure** – The data structure should be comprised of at least 3 data segments:
 - **Identity** – The unique data attributes the specifically identify the respondent including the geo demographic attributes associated with the individual.
 - **Response** – Response data from each survey conducted as reported by the respondent.
 - **Analytic** – Specific analytic data about the respondent, their response consistency and their behavior compared to group norms.

ANALYTICS

- **Base Line Stability** – The consistency of the responses to the Base Line Questions over time establishes the reliability of the Amplitude and Exploration Questions. Significant changes in the response to the Base Line Questions may indicate a material change in the views, sentiment, environment or attitude of the respondents. Expect slight variations survey to survey, however significant variations warrant further investigation.
- **Base Line Trend** – The trend of the responses to the Base Line Questions indicates the sentiment movement of the respondents. One observation does not make a trend, however several observations can give insight into the direction that the population is moving with respect to key issues. Are economic conditions causing the survey base to be more or less optimistic? Is company performance enhancing the loyalty to the organization? Identification of these trends enables a more accurate interpretation of Exploratory Questions by looking at the correlation between results.
- **Outlier Identification** – The identity of individuals and or groups existing outside the normative standards of the survey population is established. Once a normal response range is determined, the identification of Outliers permits the rapid identification of disaffected individuals or groups. These are the individuals or groups that tend to have a disproportionate influence on the balance of the population. Prompt attention to these Outliers can significantly reduce long-term remediation cost.
- **Outlier Classification** – The identification and description of the common attributes of the individuals and or groups defined as outliers. By identifying the characteristics, that the Outliers have in common an organization can assess if there is a policy, procedure or action that affects a subset of their survey population to a greater extent than the general population. If the analysis identifies a protected class as being adversely affected the organization is able to take immediate action to remediate the issue and avoid significant exposure.
- **Outlier Composition Change** – The identification and description of the attributes that have changed between the most recent surveys or over an observed time period.

A change in Outlier composition indicates the organization's ability to identify and resolve prior issues in a timely manner. If protected classes continue to be overly represented as Outliers the organization is not exhibiting pro active response to a potentially significant issue.

ACTION

- **Exclude** – Make a determination to include or exclude the response data from the Outliers in the survey results.
- **Resurvey** – Establish a method to resurvey the Outliers on a more periodic basis that the balance of the survey population. The purpose is to determine if the responses are the result of personal or environmental influences that were present at the time the respondent completed the survey that are not normally present.
- **Investigate** – If the analytics suggest that the Outlier responses are normal and consistent over an observed time period investigate what makes this individual or group unique and outside of the norm. It is also important to observe 1) whether the Outlier group population is shrinking or growing and 2) are the boundaries of the normative population converging or diverging from the Outlier responses. This provided early insight into trend movement.