Paper 128-2011

Consumer Data Analytics:
The Good, the Bad, and the Possibly Deceptive

Will Neafsey
Ford Motor Company, Dearborn, Michigan

Ted Lang
SAS, Southfield, Michigan

Abstract

Customer Intelligence relies heavily on consumer data, yet surveys are riddled with the absence of “good” data. Regardless of the imperfections, companies need to rely heavily on input from buyers to obtain a measure of the attitudes of the market.

Research firms apply various techniques in their sampling and survey methods to ensure they are collecting meaningful responses, including monetary and award offers. For various reasons, respondents, who are not engaged in the topic, still manage to find ways to get around the system to get their responses accepted to collect their survey prize. Because surveys produce a degree of bad data, we believe the analytics stage doesn’t begin after data collection; it must absolutely take place during, or immediately following, the data collection phase.

Introduction

Consumer products companies increasingly rely on survey data collected through a variety of modes including the internet, point of sale, or and face-to-face interviews. Many companies resort to "buying sample" for their surveys from large panel houses in hopes of gaining improved insight into shoppers' desires and impulses, only to be disappointed with, or misled by, the resulting data. Unfortunately, the collection method, along with the reward offers, adds another level of encouragement for respondents to provide problematic information. This paper focuses on the analytic challenges concerning consumer data by providing a practitioner's view of the analytic techniques, tests, and methodologies to help cope with the good, the bad, and the possibly deceptive aspects of survey data.1

In the realm of consumer surveys, compensation can be a double-edged sword. It plays an important role in promoting improved data quality. It provides a means to recruit a representative sample, encourages a more diverse sample, and gives an incentive to be truthful. However, it can encourage inaccurate responses since some responders become professional survey takers for the purpose of earning award points or implied better (future) treatment.

The need for better, more reliable consumer data in the auto industry spans various organizational functions. Product development and improvement, where product cycles are measured in years, depend heavily on these data. Consumer data impacts the strategic decision making regarding product mix and new markets, and is unquestionably critical in segmenting consumers in the available market space.

The analytics process depends on survey responses that are honest, complete, properly scaled and measured, and reflective of true feelings or behavior. But experience shows that getting useful or usable data is difficult. Respondents can give you a range of responses from the perfectly honest, well thought out to the blatantly false. The challenge lies in determining which is which.

There are several reasons, beyond flat-out fraud, consumers provide bad responses--long surveys, boring questions, confusing question batteries using industry-specific terminology, translations issues or culturally inappropriate questions, or lack of knowledge due to poor screening of respondents. We discuss approaches to address these concerns from simple, rules-based logic for identifying bad responses to methods that involve more advanced analytics to discover more hard to find cases. Our clear-cut objective is to screen actual bad responses while eliminating no--or as few as possible--"good" responses.

Detection of "Bad" Responses

Whether a response is good or bad is subjective as there are no known bad cases to use as targets in modeling. With the analytic techniques covered here, and considerable domain knowledge, the reliability of the data can almost certainly be improved through analytics.

The techniques we deploy to identify bad responses cover a diverse list of offenders. "Straight-liners" answer at the same response level repeatedly, or they respond in a narrow range of a response scale. Beyond that group, others answer persistently in the upper or lower part of a scale, while some simply fail to answer all the questions (particularly, questions about personal finances). In some cases, respondents simply do not understand the intent of the questions asked—either out of haste or confusion. Finally, in the case of long, reading-intensive studies, fatigue can play a large role in causing bad responses.

More flagrant examples of bad responses occur when the respondent intentionally attempts to deceive the surveyor. Some respondents try to provide duplicate submissions by representing themselves as two different people. Random responses (i.e., ABACAB answer patterns) allow respondents to appear legitimate without some enhanced examination for detecting these ne'er do wells.

Tests of Within-Respondent Concerns

Consumer surveys usually contain numerous questions covering a broad spectrum of concepts to ensure the complete capture of respondents' values, wants, and needs. In short, we ask for a lot of information, and

[1] We discuss the receipt of bad data from an analytic perspective apart from data management issues. Fixing typographical mistakes to improve cross referencing, for example, is not covered in this paper.

inattentive or scheming respondents may quickly pass through a group of questions by checking the same, or nearly the same, response level over the entire question set.

Standard Deviation Test

In our typical survey, there are enough questions within a battery to use a standard deviation test to identify those respondents with low variability in responses. These are often respondents apparently aware that straight-lining will disqualify them; so, they answer mostly at the same response levels but deviate on a small number of questions (5-5-5-5-3-5-5-5-5-6).

This simple test first calculates the standard deviation of responses within a battery for each respondent. The distribution of standard deviations of all respondents is observed, and those values in the low tail are considered for elimination.

The appropriate standard deviation cutoff is determined empirically. This is a critical point: if the cutoff is set too high, there is a risk of removing good respondents that simply use a small range on the scale because that reflects their true responses. In our recent tests of questions with a 7-point scale, the chosen cutoff has been a standard deviation value of less than 1. Of course, the number of responses captured using this test depends on many factors, including survey methods and population surveyed. The number of respondents flagged by this process can range from as little as 5% to as much as 35%. Our experience with US data shows that approximately 5-10% of our raw data are identified through application of the standard deviation test.

Duplicate Response Test

Very clever professional survey takers may actually take the same survey more than once to increase their reward. Once identified, the question remains whether to eliminate these respondents entirely from the database since the reliability of any of their responses is questionable, or to eliminate only the duplicate observation with the assessment that the other response is valid.

Tests of duplicates are typically performed by panel providers and research firms as the data are collected. Respondents entering multiple responses are getting more creative to beat deduping rules. For example, completing surveys under multiple email addresses, street addresses, and IP addresses. Identical responses and response patterns, particularly on screening questions, can indicate a duplicate respondent.

Identifying the duplicates can be a very manual process. The cost/benefits of the work must be considered before beginning the undertaking. Often, identifying duplicates can be done by comparing a small number of critical fields that it would be highly unlikely to have duplicate responses, even in a large dataset. See Figure 1.

Respondent ID	Month Born	Day Born	Year Born	Gender	State	Marital Status	Education Level	Current Occupation	Household Income
77413	September	26	1963	Male	Ohio	Married	Completed college	Owner, self-	$80,000 -
122420	September	26	1963	Male	Ohio	Married	Completed college	Owner, self-	$80,000 -
79779	February	13	1935	Male	New York	Married	Some college	Retired	$40,000 -
87183	February	13	1935	Male	New York	Married	Some college	Retired	$40,000 -
231030	January	31	1943	Female	New Jersey	Married	Completed college	Teacher,	$100,000 -
113482	January	31	1943	Female	New Jersey	Married	Completed college	Teacher, educa	$100,000 - $119

Figure 1. Duplicate survey responses

After identifying the duplicates, it is a simple process to eliminate the duplicate observation from the data.

Tests on Respondents in Comparison to Others

In this section, we look beyond the individual respondent to identify nonconforming responses. These techniques consider each respondent in relation to all others in the survey to assess whether his or her responses are kept.

Entropy Test

Savvy survey respondents avoid straight-lining in some instances by varying responses alternately using the high and low ends of the response scale. This behavior is partially addressed with the entropy test.

Entropy provides some measure of the diversity of classes used by each respondent in a battery. The entropy of responses within a battery is calculated for each respondent. The respondent's entropy is compared to the distribution of other respondents, with those values in the lowest tail being considered for elimination. This method generally catches "straight-liners" that will occasionally use a very different value (e.g. 5-5-5-1-5-1) to beat the straight-line programming. If the cut-off is set too high (inclusive), valid respondents may be removed from the sample who legitimately tend to use the extremes of the scale.

The identification of respondents can be further improved by combining of the application of the standard deviation and entropy tests. Generally speaking, the high entropy, mid-range standard respondents represent the bulk of the "good" respondents as represented in Figure 2.

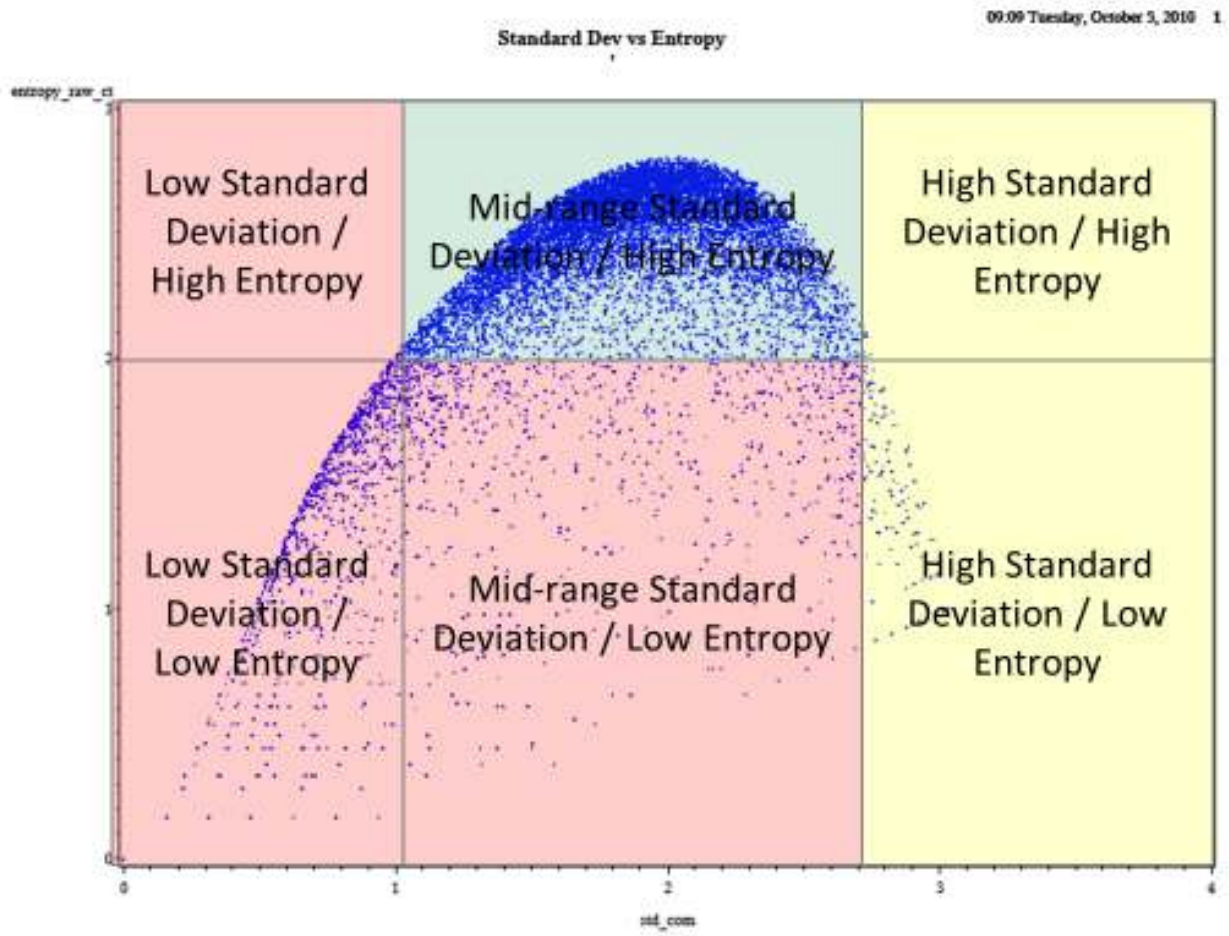


Figure 2. Graphical depiction of survey responses using Standard Deviation and Entropy tests.

Segmentation Test

In the segmentation test, clustering is applied to the data for outlier detection. Simple k-means clustering can often identify respondents that have unusual response patterns. The difficultly still lies in determining which patterns are to be classified as "bad." By working through a number of clustering solutions and profiling the results, a good portrayal of the data quality is obtained. Mapping the data using Principle Component, Factor Analysis, or Discriminant Analysis performed on question batteries assists in locating clusters with questionable data. See Figure 3 for example output where Proc CANDISC is used to develop the map of the clusters.

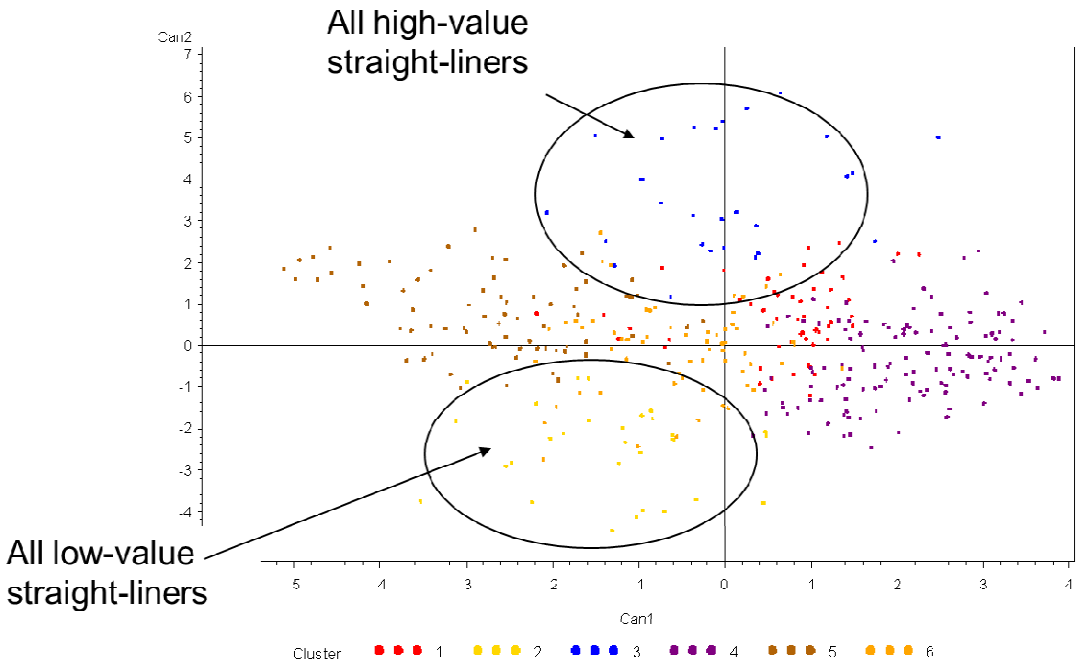


Figure 3. Map of clustering or respondents derived SAS© Proc CANDISC.

Multidimensional Graphical Techniques

In many cases, by mapping the respondents using a linear combination of many of the variables, the all of "bad" data will plot in a single location, skewing the graph. See Figure 4. Although the diagram below has the bad data colored for the sake of this paper, the skew is often clearly visible on a simple one-color scatter-plot.

Two dimensional maps can often show areas of "bad" or skewed data. Mapping is most effective when you use as many variables in the response battery as possible. Principle Component, Factor Analysis or Discriminant Analysis can be used to create good axes for the map.

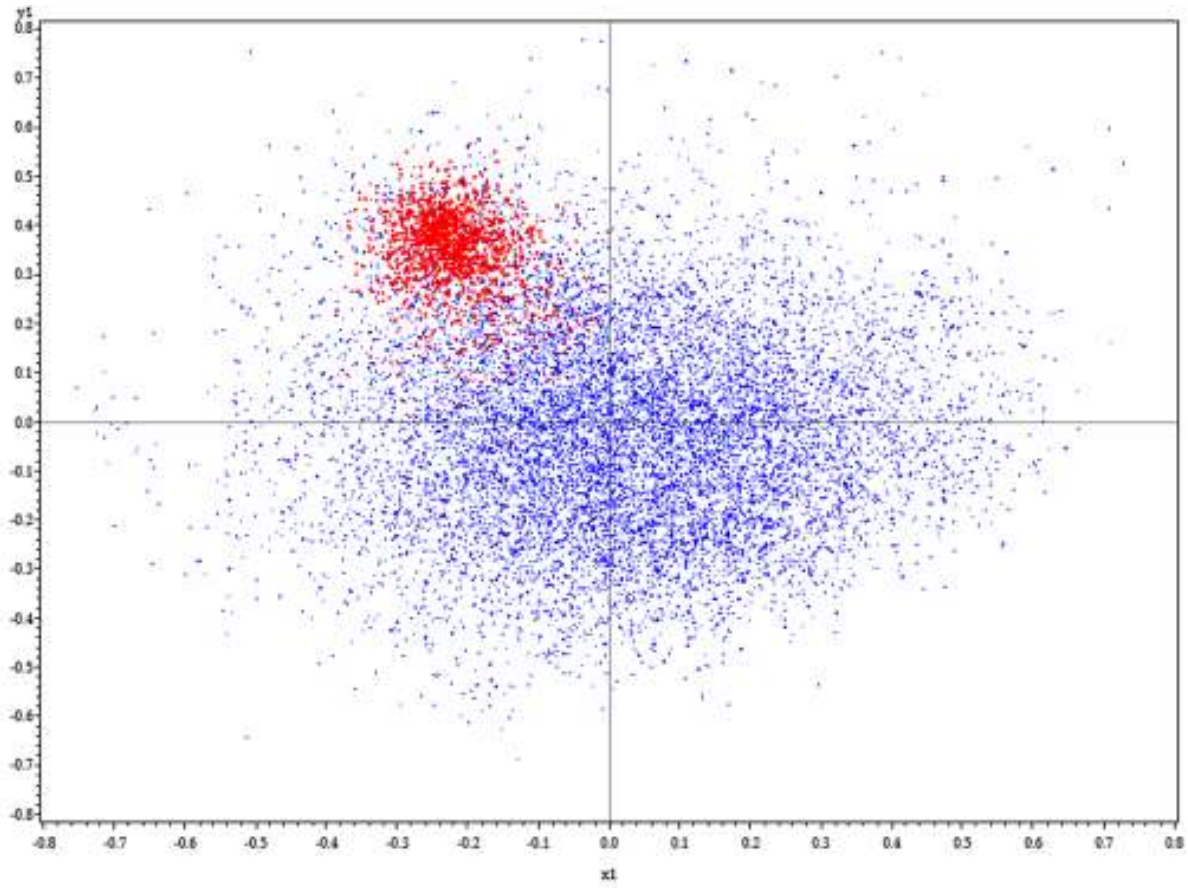


Figure 4. Graphic location of suspect survey responses

Randomness

Random survey responses represent perhaps the most challenging case of bad data because they often avoid detection using the aforementioned tests--both within respondent and across respondent tests. Respondents may alternate response levels intending to appear authentic to any machine-based detection. If they are skilled at varying their responses, the previously mentioned statistical tests will not likely identify them as nonconforming.

Random responses may still lend themselves to detection. They can lack any reasonable similarity with other respondents. For example, they may respond on the same high or low end of the scale on questions querying opposing or dissimilar concepts—this particularly helpful to identification when this behavior recurs.

Making a determination of a random response involves comparing an individual respondent to all others programmatically. Specifically, a “measure of uniqueness” is assigned to every observation. From there, some discretion is applied to determine what level of “extreme” uniqueness defines the random responses. The difficulty is that development of a "random test" is often very specific to the dataset at hand and can be extremely resource intensive from both a human standpoint and a computing standpoint.

Application of Analytics

The desired result in detecting bad survey responses is illustrated in Figure 5 showing the bad data ranking of the dataset from highest to lowest on the horizontal axis. The process requires establishing a cutoff that safely rejects the unwanted surveys and accepts the desirable ones, which requires balancing the costs and timing with the end goal of the analytic process. Recall, outside of the simple respondent-only cases, we do not have known bad responses from which to base our analysis. In the language of predictive modeling, there is no target from which

to develop models. So, we must scrutinize the data applying domain knowledge and even intuition to establish the determination of bad cases.

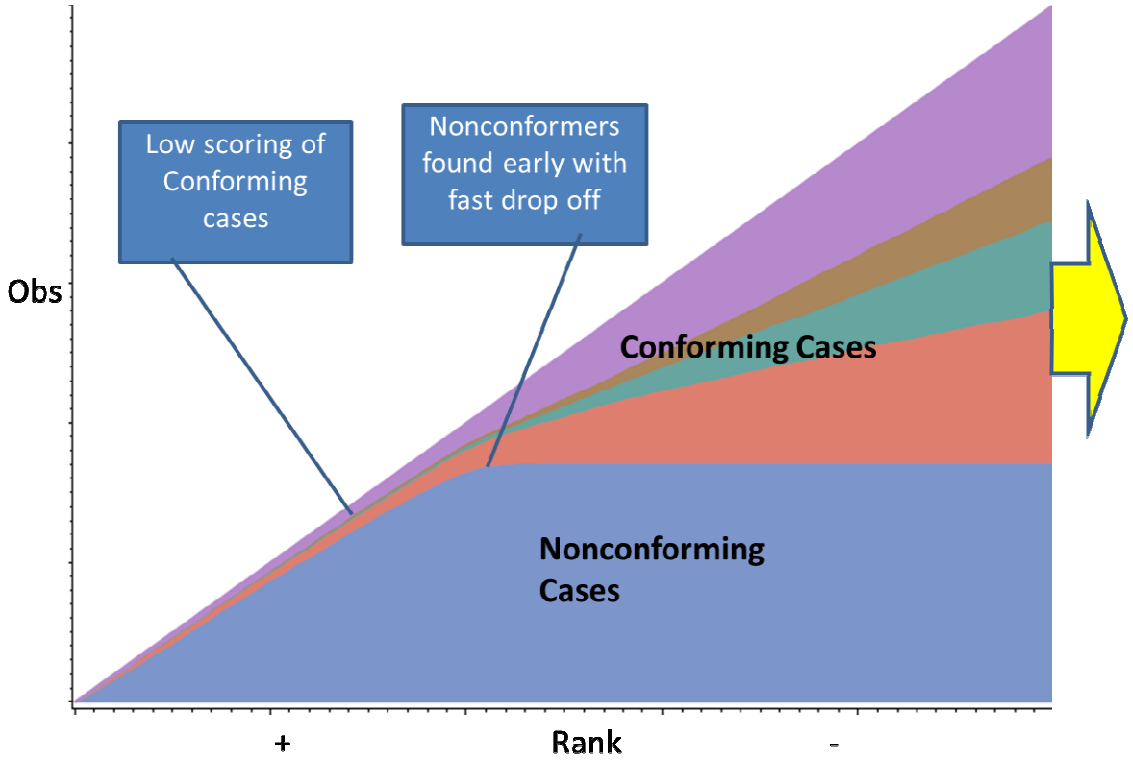


Figure 5. Desired scoring model

Fixing the Data

The simplest approach to correct the bad responses is to remove the observation from the sample. This has a cost beyond the loss of the data. For example, it disrupts the weighting or sample design, provides smaller sample sizes for analysis, and results in a loss of granularity. Finally, the deleted responses might actually be genuine responses to the survey, and the information they contain is lost.

When the cost of eliminating the suspect responses is too high, other "remedies" exists that do not require eliminating responses. The use of these techniques depends on the type of nonconformity that is involved. A survey of data cleaning approaches follows.

Imputation

In the event of missing data, respondent data can be saved by imputing the missing responses using the available responses to determine the corrected levels. This allows the use of the observations to continue analytic approaches that require a full set of answers and preserves valuable sample. However, imputation creates data where none exists.

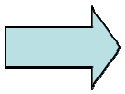
Transformation

Transforming the data is the most common solution for non-critical data issues. This allows working with data where the respondents have used the scales differently. We feel this is an acceptable solution for improving suboptimal data, but it is not without hazards. Transforming the data causes a loss some of the original information regarding strength and location of the respondent's true midpoint.

Rank

The transformation into ranks helps remove the respondent scaling issues such as using only a narrow range of the scale or answering persistently in the ends of the scale. See Figure 6.

	Q1	Q2	Q3	Q4	Q5	Q6
A	5	5	5	5	4	5
B	3	5	4	1	1	2
C	1	2	2	2	1	1
D	3	5	3	4	3	4
E	1	2	1	1	1	1
F	4	5	3	3	3	5
G	5	5	1	1	1	5
H	2	5	1	4	1	5
I	1	2	2	1	1	1
J	3	5	3	2	2	2



	Q1	Q2	Q3	Q4	Q5	Q6
A	1	1	1	1	6	1
B	3	1	2	5	5	4
C	4	1	1	1	4	4
D	4	1	4	2	4	2
E	2	1	2	2	2	2
F	3	1	4	4	4	1
G	1	1	4	4	4	1
H	4	1	5	3	5	1
I	3	1	1	3	3	3
J	2	1	2	4	4	4

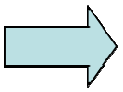
Figure 6. Rank Transformation

Z-Scores

Z-Scores preserve the relative position of each response within a battery for each respondent, but eliminate the scaling issues across the respondents.

The transformation into Z-scores removes the respondent scaling issues the same as rank transformation discussed above. See Figure 7.

	Q1	Q2	Q3	Q4	Q5	Q6
A	5	5	5	5	4	5
B	3	5	4	1	1	2
C	1	2	2	2	1	1
D	3	5	3	4	3	4
E	1	2	1	1	1	1
F	4	5	3	3	3	5
G	5	5	1	1	1	5
H	2	5	1	4	1	5
I	1	2	2	1	1	1
J	3	5	3	2	2	2

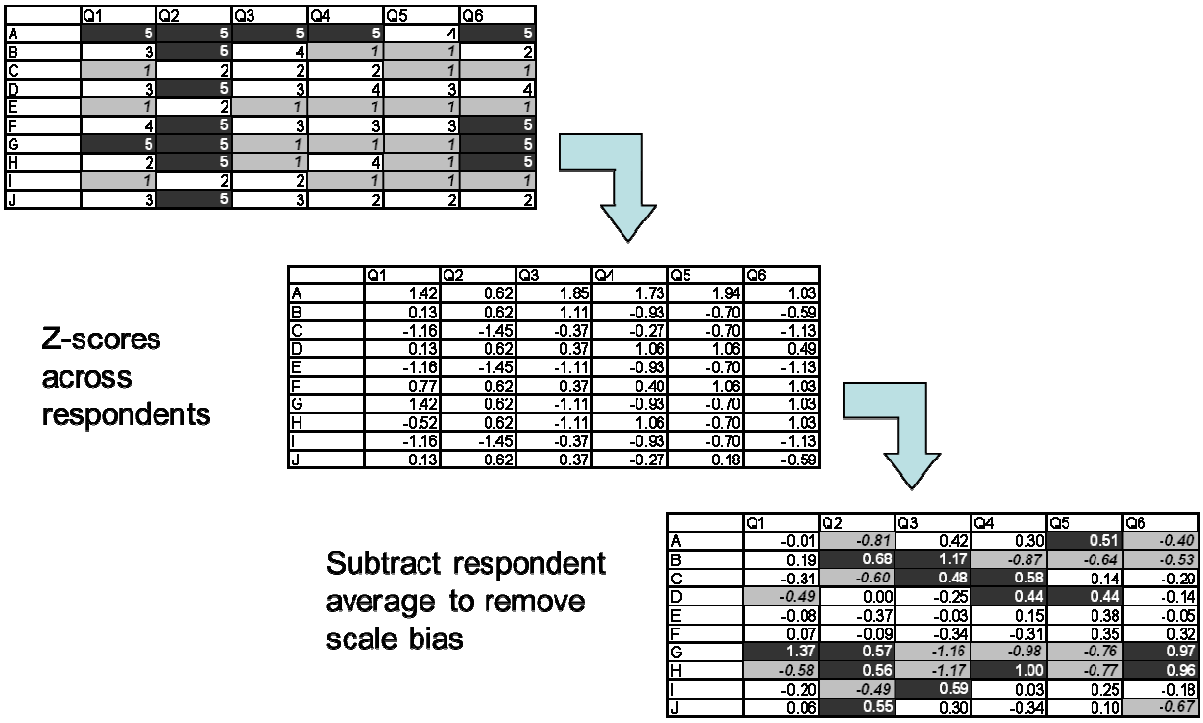


	Q1	Q2	Q3	Q4	Q5	Q6
A	0.41	0.41	0.41	0.41	-2.04	0.41
B	0.20	1.43	0.82	-1.02	-1.02	-0.41
C	-0.91	0.91	0.91	0.91	-0.91	-0.91
D	-0.82	1.63	-0.82	0.41	-0.82	0.41
E	-0.41	2.04	-0.41	-0.41	-0.41	-0.41
F	0.17	1.19	-0.85	-0.85	-0.85	1.19
G	0.91	0.91	-0.91	-0.91	-0.91	0.91
H	-0.53	1.05	-1.05	0.53	-1.05	1.05
I	-0.65	1.29	1.29	-0.65	-0.65	-0.65
J	0.14	1.85	0.14	-0.71	-0.71	-0.71

Figure 7. Z-Score Transformation

Double-Centering

Double center standardizes the data in two directions (within respondent and across respondent) to highlight the peaks and valleys of each respondent-question combination. See Figure 8.



	Q1	Q2	Q3	Q4	Q5	Q6
A	5	5	5	5	4	5
B	3	5	4	1	1	2
C	1	2	2	2	1	1
D	3	5	3	4	3	4
E	1	2	1	1	1	1
F	4	5	3	3	3	5
G	5	5	1	1	1	5
H	2	5	1	4	1	5
I	1	2	2	1	1	1
J	3	5	3	2	2	2

	Q1	Q2	Q3	Q4	Q5	Q6
A	1.42	0.62	1.85	1.73	1.94	1.03
B	0.13	0.62	1.11	-0.93	-0.70	-0.59
C	-1.16	-1.45	-0.37	-0.27	-0.70	-1.13
D	0.13	0.62	0.37	1.06	1.06	0.49
E	-1.16	-1.45	-1.11	-0.93	-0.70	-1.13
F	0.77	0.62	0.37	0.40	1.06	1.03
G	1.42	0.62	-1.11	-0.93	-0.70	1.03
H	-0.52	0.62	-1.11	1.06	-0.70	1.03
I	-1.16	-1.45	-0.37	-0.93	-0.70	-1.13
J	0.13	0.62	0.37	-0.27	0.18	-0.59

	Q1	Q2	Q3	Q4	Q5	Q6
A	-0.01	-0.81	0.42	0.30	0.51	-0.40
B	0.19	0.68	1.17	-0.87	-0.64	-0.53
C	-0.31	-0.60	0.48	0.58	0.14	-0.20
D	-0.49	0.00	-0.25	0.44	0.44	-0.14
E	-0.08	-0.37	-0.03	0.15	0.38	-0.05
F	0.07	-0.09	-0.34	-0.31	0.35	0.32
G	1.37	0.57	-1.15	-0.98	-0.76	0.97
H	-0.58	0.56	-1.17	1.00	-0.77	0.96
I	-0.20	-0.49	0.59	0.03	0.25	-0.18
J	0.06	0.55	0.30	-0.34	0.10	-0.67

Figure 8. Double Center Transformation

Summary

In an ideal world, the research firm employed to perform the survey process would only provide good responses. Short of that, wherever feasible, we provide them with some of our techniques for application at the time of the survey to the degree practicable. Regardless, because bad responses can still be found in survey data delivered from research vendors, we continue, after the receipt of the data, to apply the techniques discussed in this paper to improve the data. It costly and painful, but cleaning the survey data prior to the ultimate analysis is critical.

Although there are not, and cannot be, any clear cut rules to follow, our experience provides us with some simple guidelines for survey development:

- Create an understandable, interesting and engaging survey
  - Test it on people outside your industry
  - Change topics frequently
- Be conscious of length of time
  - Bored causes bad responses in the best people
- Create “trap” questions
- Design your battery of questions for diversity
  - One-sided responses will tip you off to bad responses
- Use ranking questions when possible
- Test your data early and often
  - Preliminary data dumps from vendors can often head off issues early while study is still fielding.
- Don’t overcook you data
  - If you have to do too much transformation, are you sure it’s still valid?

At the end of the day, if you have data, you have bad data. The real questions are, how much and where? The pursuit of bad data is thoughtful balance of removing clearly harmful data and preserving expensive and valuable consumer input. The key is to understand that after applying the techniques discussed above, the data must still be good enough to stand, untransformed, in the final analysis and report.

Authors

Will Neafsey
Manager, Brand DNA and Consumer Segmentation
Ford Motor Company
wneafsey@ford.com
www.ford.com

Ted Lang
Technical Account Manager
SAS Institute Inc.
ted.lang@sas.com
www.sas.com

Acknoledgements

Greg Brown, Ph.D.
Independent Consultant
Gregory.Brown48b@verizon.net

Ann Chen
Rightside Enterprises, Inc.
ann.chen@rightsideenterprises.com
www.rightsideenterprises.com

Sara Preston
Manager, Customer Analytics
Ford Motor Company
spresto1@ford.com
www.ford.com

Kim Vincent
Technical Account Manager
SAS Institute Inc.
kim.vincent@sas.com
www.sas.com