"BAD" RESPONDENTS AND THE PANEL QUALITY INDEX

As online research has become the dominant mode of research sample quality issues have finally reached their appropriate importance. We have developed an extensive database of 20 online U.S. and over 70 global data sources using a standard questionnaire instrument. This has allowed us to explore techniques for monitoring and potentially controlling data-source quality. This paper outlines practical procedures to feret out survey data that represents hard core quality issues..

In online research where we function in a non probabilistic environment it is at times difficult to identify and define problematic respondents. Much discussion has been centered on respondent behaviors that are considered undesirable, speeders, professionals, and satisficers have all been given considerable space in the market research literature. We propose here a method of blending these behaviors into a single index to assist researchers in rejecting the worst of those respondents. These respondents provide us with information that does not reflect the balance of the data collected and thus represents a quality challenge. There will be many steps toward online quality standards now that the industry has seemingly woken up but for now we will have to arrive at strategies that identify behaviors that bias our data as some have contended, by misrepresentation, by duplication when respondents are rewarded for participation, or by inattentiveness.

Identifying let alone predicting "bad" respondents is problematic. All responses from surveys are expected to be distributed. This is inherent in survey and opinion research. If variation of values were not expected, then it would serve little purpose to conduct them at all. However, this introduces a fundamental problem of identifying which responses are flawed.

We can screen out individuals that do not meet survey participation criteria prior to the taking the survey. This type of prior screening is a growing industry procedure whose importance is on the rise. Respondent duplicates can be removed from the results using monitoring software in online surveys[1] or simply tagged for future consideration. But for the acceptable respondents, who meet these criteria, finding a single metric that signals a "bad" respondent is infeasible. What we can do, however, is identify a number of "troubling" responses. But, no single measure will identify by itself only flawed respondents.

If a number of measures of "troubling" behavior are treated as screens, removing all respondents that failed any criteria, a major fraction of the total sample is likely to be removed. As we will see later, this might be as much as 64% of the respondents[2]. This may exclude almost all of the "bad" respondents but also will remove many more of fully acceptable ones. More troubling, however, this process would most likely produce a

---

[1] The technique of Fingerprinting is used to screen out duplicates. This involves measuring the characteristics of the browser source for the respond to a questionnaire. Software then checks to see if this is a duplicate.

[2] In a recent, white paper by MarketTools indicated that almost 25% of their panel did not qualify under their screen criteria approach and were considered fraudulent, [Michael Conklin, *What Impact Do "Bad" Respondents have on Business Decisions*, published by MarketTools (2009)]. This is probably highly unlikely for all of them to be fraudulent.

biased sample. That is, we would have thrown out the baby with the wash-water. The integrity of the sample requires that its diversity be maintained. By being over critical, we could only expect to have reduced this diversity and produce results that would not reflect the true population.

In a recent white paper, Melanie Courtright and Denise Brien[3] have proposed the use of an exclusion decision rule based on a number of measures of "troubling" behavior. Their rule involved removal of respondents identified as having violated somewhat more than half of their criteria. The procedure that is proposed here follows that line of thought of excluding respondents who have violated 3 or more out of six criteria.

METRICS OF "TROUBLING" RESPONSES

There are two types of "troubling" behavior that need to be examined: errors in execution, and aberrant behavior. Errors in execution are responses that are shown to be wrong. They either reflect confusion by the respondent or inattentiveness. They are obtained by specific questions with the questionnaire. Specifically here we are referring to two types of failure: (1) to follow instructions and (2) inconsistent responses. The inconsistent responses are based on paired questions in reverse order so that one could not consistently rate both high or both low. Either would be judged as an inconsistent response. The standard questionnaire we have employed in our research on research has provided one measure of failure to follow instructions and two measures of inconsistency.

Aberrant respondent behavior consists of a number of characteristics that while they are not inherently wrong as in the case of the errors in execution, they are possible indicators of potential problems. These include professionalism, speeders, and straight-liners. Professionals here merely refer to frequent survey takers or those individuals belonging to a large number of online panels. This measure is captured by specific questions in the questionnaire.

Speeders are those respondents that complete the survey in a very short time. It is believed that this indicates inattentiveness in execution or at least the potential for less than fully considered activities. The lowest ten percentile point[4] was chosen for the transition point defining speeders, as shown below, Figure 1.

[3] M. Courtright and D. Brien *The Devil is in the Data* published by DMS (2009)
[4] The ten percentile point also represents a deviation from the exponential portion of the distribution.

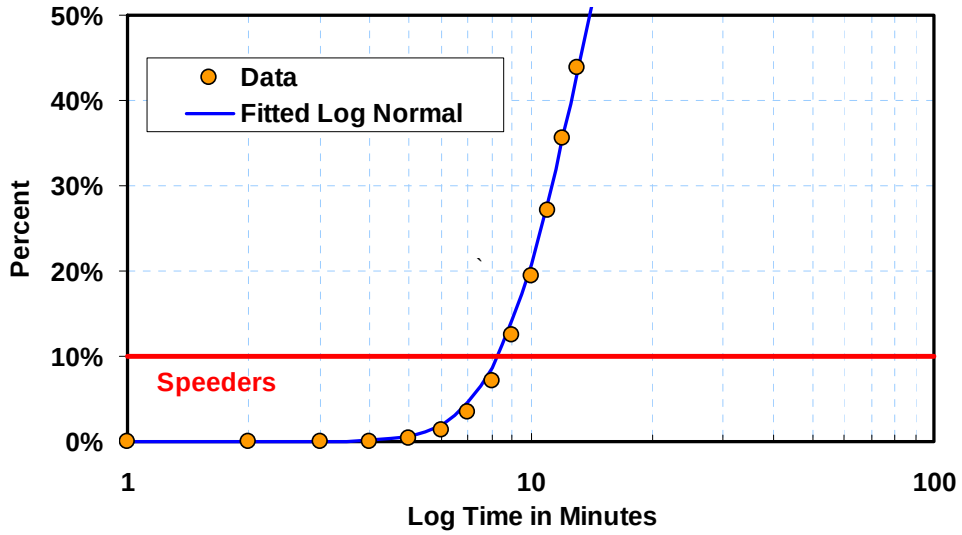

Figure 1, Defining Speeders, Distribution of Execution Times

Straight-liners are survey participants who have little variation in their responses. There are some respondents who have no variation. But many more, however, indicate only very little variation. That variation is computed as the standard error around a set of similar questions. In the case of the standard questionnaire, this was based on 31 opinion questions. The standard errors for respondents varied between effectively zero to over three on a seven point scale. Having a standard error below 1 was used to define straight-liners[5], as shown in Figure 2.

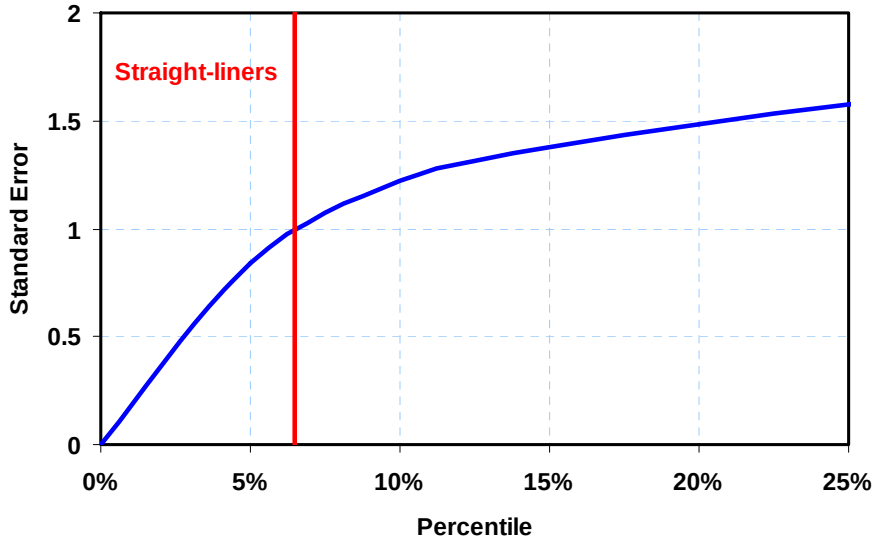

Figure 2, Defining Straight-liners, Distribution of Standard Errors

[5] The point appeared to separate two almost linear portions of the distribution.

THE QUALITY INDEX AND SEGMENTS

A quality index was defined for each respondent by the number of errors and indicators of aberrant behavior. The index goes from zero for those without error to six for those who appear to get everything wrong. These can then be grouped in quality segments. Following the convention by Courtright and Brien[3] four segments are defined with those with 3 or more errors designated as the "Worst" segment. Those without error are considered "Ideal" with one error designated "Typical" and those with two errors as "Imperfect". The results from our database are shown on Figure 3.

These assignments could be merely random. That is, error and aberrant behavior could be just an event and not associated with other expected "bad" behavior. On Figure 3 is also the expected distribution if these errors were purely random events. Note that for the first two segments, the frequencies are the same or at least within statistical precision. Not so for the last two. The frequencies are significantly different indicating that this is not simply a single independent random process, but are in some way interconnected.

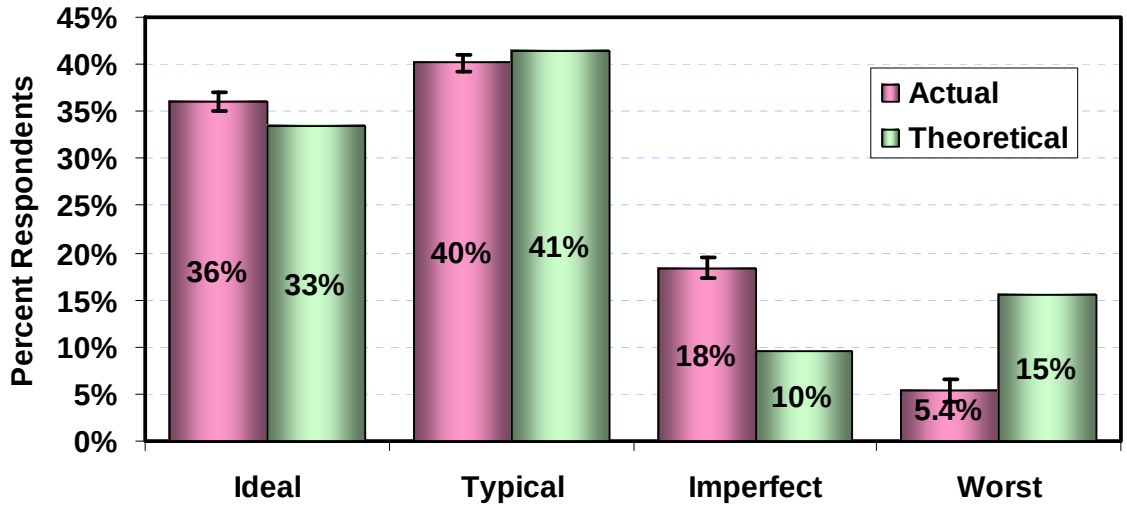

Figure 3, Frequency of Quality Segments

We should expect that the "bad" respondents would show distinctly different characteristics than the other segments. This is shown in terms of average values across questions in the survey. Figure 4, shows this result. Note that there is only a small, statistically insignificant difference among the three lower error segments while the "Worst" segment stands out, showing a significantly different value[6].

[6] Courtright and Brien[3] show this distinction among quality segments in much more detail in terms of specific respondent behavior as well as attitudes.

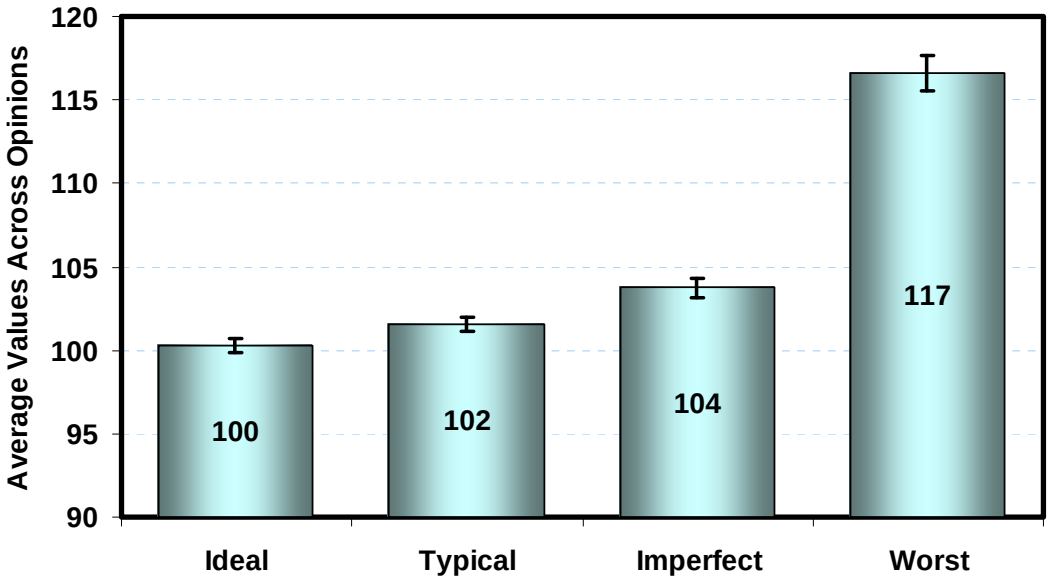

Figure 4, Average Values of Opinion by Quality Segments

VARIATION ACROSS DATA-SOURCES

This identification of "bad" respondents can, of course, be used to improve results by excluding them from survey sample. This is a relatively average small frequency and their exclusion should not affect the integrity of the rest of the panel. However, that frequency can differ widely among panels. In Figure 5, the distribution of quality segments is shown across 17 United States online panels and data sources. These differ in respect to most of the segments. It should be noted, that these panels have vastly different structures, management conditions, and respondent sources. It is, therefore, not surprising that they would also differ in their incidence of errors.

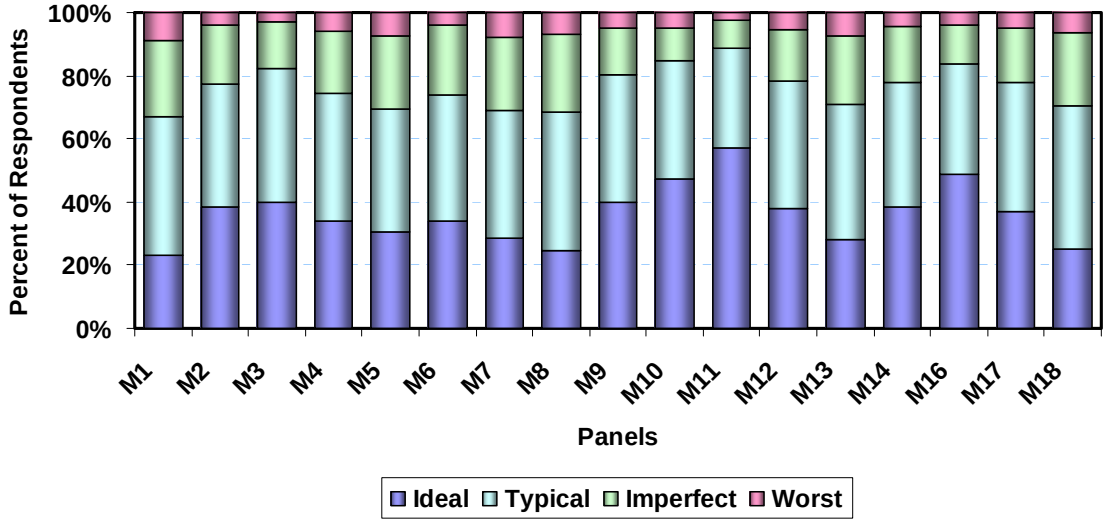

Figure 5, Distribution of Quality Segments by U.S. Data Source

While the distribution is interesting, the focus should be on the "Worst" segment. This is the segment containing the respondents that we might wish to exclude. In this respect, it

is a composite measure of panel quality. This is shown on Figure 6. Note that the frequencies vary over almost a factor of four with some almost twice as high as the average value.

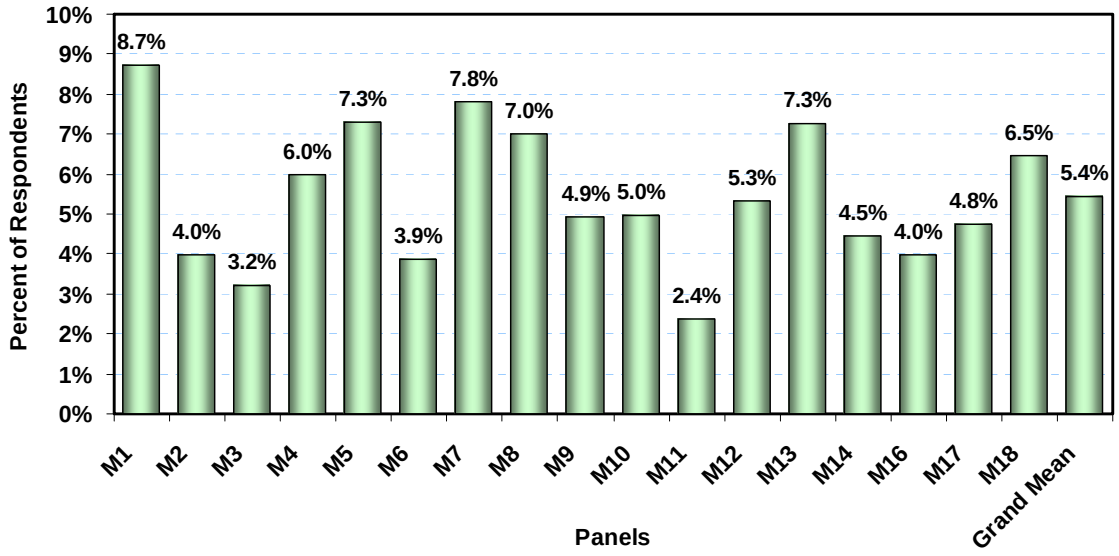

Figure 6, Frequency of Worst Segment ("Bad" Respondents) by U.S. Data Source

Practical Issues:

Market research practitioners have gravitated to the web for its obvious advantages of speed and cost. Quality issues lagged until about three years ago and have now become an important driver. Elimination of problem respondents has been a piecemeal practice usually relying upon the removal of one designated group or the next. Combinations of behaviors have rarely been entertained.

The cost of keeping respondents who contribute questionable data is incalculable. However the cost of removing them is clearly measurable and consequential. For some there is an undercurrent fear that costs will rise beyond the tolerance of the market to pay for them, or even worse, competitive position will be lost in today's bidding environment.

The measures suggested here are in our opinion minimal. We doubt that quality respondent sources will penalize researchers who employ a method which on average forces replacement of 5.4% of all survey takers.

We are confronted by a cultural shift in our research. Those who use our data to make business decisions cannot afford spurious conclusions to driven by dubious respondents. Yet there is a consistent fear that end users will refuse to pay the tariff if one exists.

Here the greatest cost lies in the questionnaire real estate that the quality metric questions require. Measuring speeders and straight liners is nominal in that it is a programming function that is not represented by questionnaire wording although there are some design

considerations. However, the four questions that measure the two inconsistencies, one that captures the frequency of survey taking and another that serves as a “trap” could amount to about ninety seconds. The extra length has a direct cost however nominal.

Conclusion:

It is past time to coordinate quality metrics. Here we provide a metric that requires six measures. The differences between panel sources should alarm researchers sufficiently to employ techniques of this sort.

Combined with digital fingerprinting to remove duplicate respondents these behavioral measures are a beginning. One might think of them as a sliding scale. We have set the bar at three failures and leave it to other to intensify the standard. For now, we hope that practices such as this become standard practice.

As we continue to grow our database of respondents around the world we will be able to ground out quality metrics to a larger scope. For now the data in the United States has provided us with this launching point. We see no reason that similar methods could not apply in all global markets.