

Consistency in Global Non-Probabilistic Online Samples

Steve H. Gittelman, Ph.D. and Elaine Trimarchi, Mktg, Inc.

Researchers must know if the changes they see in their data are real or artifacts of an inconsistent sample frame. The quality of respondents as measured by their levels of engagement is hypothesized to be correlated with the consistency of data obtained from commercial online panels around the world. To that end, identical tracking studies were conducted among twelve panel companies in six countries (United States, United Kingdom, Canada, Brazil, India, Russia and China). A correlation was found to exist between respondent disengagement and the inability of a panel to generate replicable, consistent data using the same survey vehicle.

Consistency in Global Non-Probabilistic Online Samples

Consistency of online samples has become a core issue for market researchers. After all, much of the value we provide is in the tracking studies we perform, but even one-time studies should relate to some reference and not float in a sea of variability between panels. If your data changes, it is essential to know if the changes are real or the inadvertent product of sample inconsistency.

In the past, we had no reason to fret over sample consistency. At the core of every research career there is a fundamental reliance on probability. Toss a coin, any coin. It will reliably come up half heads and half tails. There is no magic in it; in the coin toss exercise we are matching to known characteristics.

No one would expect to toss a coin a million times in order to prove its “fairness.” Market researchers have drawn samples from known commodities for decades always relying on the fairness of a coin toss. Households could reliably be reached by telephone almost 99% of the time and the small fraction of non-phone homes mattered little. Yes, we had to adhere to strict calling regimens, callbacks, refusal conversions and most of all recovering a large percentage of the sample. The key here is that the telephone sample replicated the census because it reached most segments of the population with equal penetration and theoretically equal probability. We knew the composition of the population in advance and could rely on that reference to keep our samples targeted. It was reliable, predictable and repeatable and thus consistent; it earned its name “probability sample”.

As refusal rates began to climb and “do not call lists” became good politics, the ability to reach some segments of the population dwindled. The all-important concept of a high recovery rate began to crumble. Phone, with an increasing percentage of line-cutters and cell phone users, had begun to dwindle, as its ability to replicate the census was impaired.

Why worry about the census now?

Where once research was well grounded in a probabilistic framework supported by an underlying census of the population, online market research has moved into a new era, from a probabilistic framework to “working without a net.” In the absence of a probabilistic method to anchor samples, non-probabilistic samples can drift without our knowing.

One, now historic, example of this happening was presented by Ron Gailey (IIR 2008), now of Coca Cola, previously of Washington Mutual, who disclosed how 29 studies representing 40,000 online interviews had gone astray due to panel inconsistency. In the WaMu research, the change was due to shifts in respondent tenure that resulted from changes in the panel’s constituents over the two-year span of the base research. Gailey’s research showed a 30% drop in buyer demand for WaMu’s financial products; a result (2006-7) unsupported by sales. His conclusion, after much study, was that long-term panel members were less optimistic about their purchases than new panel members. Others have since corroborated this finding. The lingering question, now that WaMu is gone, is how the tainted research impacted on critical business decisions.

The effects of hyperactive respondents and other online respondent ills were brushed under the rug. Ron Gailey unknowingly had to use a sample that showed aging affects that took time to evolve.

Ron Gailey had to do a lot of digging to find the root problem within his data. If online samples changed as they aged, then they could not be counted on to provide reliable data through time. And there are a host of reasons that could change them. For example, mergers bring together samples of different sourcing and aging profiles. Management makes decisions influencing the frequency of hyperactive respondents by increasing the number of surveys that they are invited

to and allowed to complete. It is evident that a panel that is used many times a month is different from one that has new respondents all the time. Panels differ for a wide variety of reasons, many of which are not disclosed to clients since currently there are no standards.

Luckily, there is a world out there of science that has long ago learned to collect data and make decisions based upon sampling frames that are non-probabilistic. When Charles Darwin hauled himself onto the volcanic shores of the Galapagos Islands he took samples of as many islands as he could reach. For the most part, these isolated little islands were different from one another. Even birds that could theoretically fly from one to the next differed. He didn't have a census to draw his conclusions: He was the census!

Charles took samples of a few islands and wrote a pretty good book. The samples were not grounded in probability theory and he could not generalize from island to island. Vive la difference! It was the differences that gave him clues. Each island was an ecosystem unto itself and the differences that species on the islands had to endure shaped them into the specialists that they became.

Our use of online data has much to learn from island biogeography. Think of each online panel as an island. They have similarities but are drawn from different sources. We should not expect them to be identical; we should expect them to be different. Our research has shown them to be quite inconsistent (Gittelman and Trimarchi, Feb 2009 CASRO) and the ARF supports this point. The panels are not interchangeable. The online panels are drawn from different sources, are subject to differing management practices and for a host of reasons yield different results.

Hidden in all of this, is the concept of sample consistency. After all, if we measure bias and can't anticipate its shifts over time, then we will not understand which changes are coming from our

data or from background noise in the sample. Thus, as the ARF announced in June of 2009, the issue of consistency is the most important area of concern. We must learn to measure not only what the constituent elements of our data sources are, but also how they change over time. In other words, we have to enter a new world of replicable method for consistency analysis.

As we do our research, we must know what the changes in our data mean. Are they the product of shifts in opinions or changes in the sampling frame? To get a grip on this we need parallel studies that document the consistency of our samples.

Consistency is a complex concept. We need to know the differences between panels at any given moment so that we can understand how the panel we use changes through time and events, confidently switch to a different panel, or use multiple panels in our research. Blending samples is a good way of spreading risk among many to avoid the potential ills of just one. Be prepared for the use of sample blending techniques to become an industry standard for achieving consistency around the globe.

We have moved onward from our initial tracking study of the American markets (Gittelman and Trimarchi, 2009) and have expanded our research to include over 200 panels in 35 nations. In each, a standard translated instrument was used, that included a diversity of measures but mostly focuses on buying behavior segmentations. By conducting repeated waves of this consistency tracking study, a local *Grand Mean* was calculated for each market. In addition, using standard quality control techniques, an analysis of the consistency of each panel is conducted.

It is quite predictable that buying behavior will shift on a seasonal basis. It is also predictable that the ice will melt in spring and that the rains will come. We bet on it all the time. They say that the only thing that you can count on is death and taxes...wrong! Predictability is as much a part

of consistency as is reliability. Consistency does not mean staying the same, but rather having predictable patterns of change. The U.S. census did not provide us with that.

If all panels were required by their users to show that they were consistent, we would have both a measure of quality in their samples as well as a new set of indices to replace the absence of a usable online census. Certainly, if all panels provided data on how their members responded to a battery of purchasing questions and the segmentations were tracked, Ron Gailey would have had a reference to consider when his panel's composition began to change. The census has very little relationship to the online community, and even less to the online community that participate in panels. The offline population is different from those online and the panels themselves are a disparate group with almost no guiding standards. The whole process of belonging to a panel filters out an unknown sector of the population and no one knows how to weight them or what problematic and un-weightable variables might be hidden in the data.

We test panels regularly for consistency: each participates in at least three waves of audit separated by a minimum of three months. This provides end users with assurances regarding the stability of panel output.

In this paper, we summarize the results of this consistency analysis conducted in six countries, the United States, Canada, Brazil, Russia, India and China. Consistent data is either the product of real changes in the underlying phenomena that are being measured or are driven by the data collection method. The former is clearly not a problem while the latter clearly is.

Here we use demographically balanced populations (age, gender, income) from wave to wave, an identical questionnaire instrument and track data changes within a single panel. We use two

methods of data collection, one where the cooperating panel is informed of our efforts and the second where the sample source is blind.

We divide our six target countries into those where online penetration is high and online research is well established (Canada and the United States) and the “BRIC” countries (Brazil, Russia, India and China) where online penetration is low and online research is a relatively new endeavor.

The identity of participating companies has been kept confidential.

Methods

Question Types

Analysis was based on response to questions on three types of information:

- ***Demographics*** (including age, income, education, and marital status distribution) reflect the traditional classification of respondents. Note that some of these variables were used to quota control the sample. As such, they measure the consistency of the quota process.
- ***Structural Segments*** (based on buyer behavior, sociographic issues, and media use) reflect the cultural, social, and behavioral characteristics of the respondents. These segmentation schemes may vary between countries and regions. They should be more consistent within countries and within panels over time.
- ***Source Performance*** (including erroneous, professional and satisficing behavior metrics) may reflect the quality of survey results. These include the issues of incorrect responses, speeding through the survey, and frequently participating in surveys.

Two references are used to gauge the consistency of results: overall average response (local mean), and the external “Grand Mean” representing the average responses for a standard questionnaire over a number of sources. The Grand Mean references have been collected by country and are used only within the country to evaluate panels.

Consistency is evaluated as a comparison with the average values for a given panel across all waves of the consistency analysis. Large deviations from the average are labeled as ‘inconsistencies’. The *Grand Mean* exists purely as a point of reference to our estimate of a country’s ‘actual’ population value. The reliability of this estimate varies according to the amount of data and diversity of sources collected within that country.

References and Error Bounds

The internal reference for analysis is based on a moving average of the data series. As more data in the series is available, the average reference values are expected to become increasingly stable. All variations are assumed to be associated only with the tested panel. References are treated as population values, with no error bounds. Eliminating this source of error results in a decrease in the error bound making this analysis somewhat more conservative.

Distribution Metrics

Differences between values within the data are tested and depicted using the following methods:

- ***Variation & Error*** – Stacked bar charts are used to show the time series results of the sample-set along with the appropriate references. Error bounds at 2 standard errors around the

components are also shown to illustrate relative importance of differences. As previously noted, all error is assumed to be associated with the panels being examined.

Distance Measure of Variation – The Root-Mean-Square distance¹ measure is used to indicate the degree of separation between two samples. In a broad sense, it can be approximated as the ‘average’ percent deviation from the panel average on a given metric.

BEHAVIORAL SEGMENT DISTRIBUTIONS

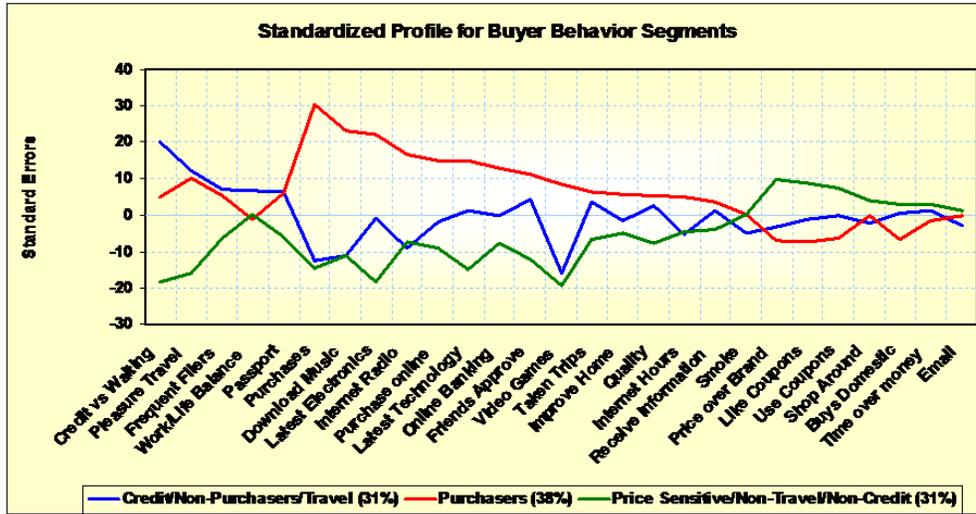
Typically, panels and lists are filtered to balance demographics against some external standard such as the known general population. However, demographics do not account for every attribute of importance to a market researcher. In an attempt to account for the behavioral attributes of respondents as well, we introduced three segmentation schemes, determined through cluster analysis on an array of 62 consumer behavioral variables.

Three segmentation schemes are being used in this evaluation focusing on: (1) Buyer Behavior, (2) Sociographic Factors, and (3) Media Use Factors.

BUYER BEHAVIOR SEGMENTS

The buyer behavior segments are intended to capture the variability in the attitudes and actions regarding the purchase of a broad range of products. The standardized profiles for the selected US sources are shown below and reflect the response to 37 input variables.

¹ The root-mean-square distance is defined as the square root of the average of square of the differences between the distributions elements.



The titles of the segments reflect the strongest loading variables making up the segment. The purpose of this scheme is to reflect differences between sources of data and the general Grand Mean representing that region. The distribution of these segments can vary widely between different countries and global regions. These are expected cultural variations. However, we expect the distribution of these segments among panel and sources of data within regions to be less variable. Furthermore, the distribution of segments should be consistent over time within a panel or source, assuming sampling methods remain constant.

SOURCE PERFORMANCE METRICS

Aside from the behavioral profiles of our respondents, their survey-taking behaviors are also of interest to us. These extend to questions of both quality (or engagement) and of hyperactivity.

Hyperactivity - Hyperactivity reflects the frequency of survey-taking by the participants in the panels. In general, these focus on issues and concerns with the long term maintenance and in particular the tendency of containing “professional” participants. These metrics may include participants who belong to multiple panels, have been on panels for an extended period of time or who take surveys frequently.

Trap Questions- The incidence of errors in the execution of questionnaires reflects the quality of the panel. These are “checks” designed into the testing instrument. They include but are not limited by: (1) inconsistency in responding to multiple questions and (2) the failure to follow instructions.

Satisficing – Some respondents can exhibit behaviors that put their responses in doubt without failing trap questions. These include: (1) “speeders” who finish their questionnaire in extraordinarily short time and (2) “straight-liners” who tend to give the same answer to a large number of questions. These are not errors, just extreme behavior which provides a warning of potential problems.

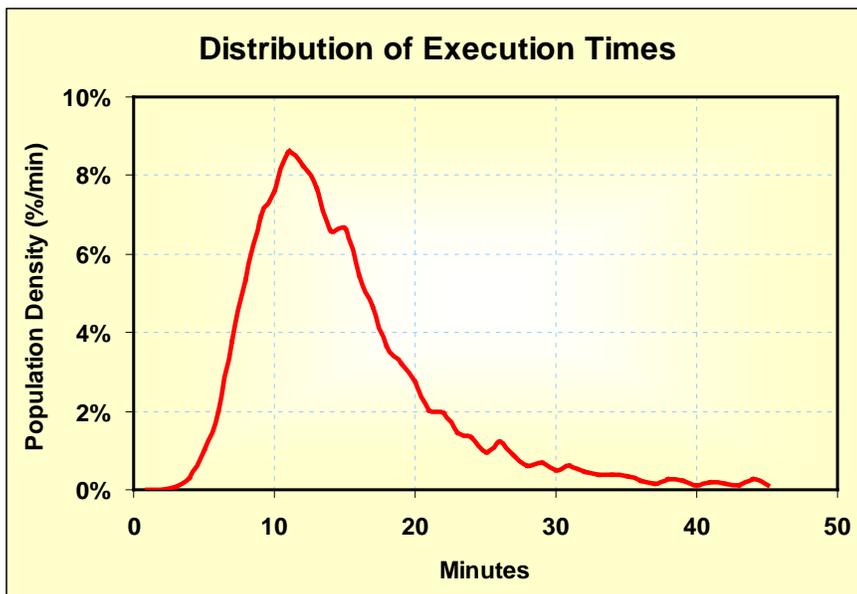
HYPERACTIVITY (FREQUENT SURVEY TAKERS)

A concern regarding online panels is the development of “professional” survey takers. These are members of the panel or data sources that are frequent survey-takers. The frequency of professionals is estimated in terms of four measures: (1) belonging to 5 or more panels, (2)

taking surveys almost every day, (3) having taken at least 30 surveys in the past month, and (4) panel tenure. All of these are self assessments and as such may be in error, but they represent consistent metrics.

SATISFICING BEHAVIOR (SPEEDERS AND STRAIGHT-LINERS)

Below is a typical distribution of completion times for the test survey. Note that it has been truncated at 45 minutes. Because of the nature of online surveys, participants may delay execution and thereby run up huge apparent elapsed times. Typically these long times are removed for analysis. Speeders are those that finish the questionnaire very fast. Generally for our test instrument, that is less than an overall lower 10 percentile.



Similarly, straight-liners are defined based on the lack of variation in their responses. For the purposes of analysis, the standard deviation over a range of similar questions is used to estimate variation. Straight-liners are defined for this analysis as respondents with a selected standard deviation of 1 unit (out of 7) or less for 30 questions.

Q-METRICS®

The Quality Segments are based on the number of faults recorded on the above metrics including errors, “professional” behavior, or satisficing. There are six indicators in this estimation: three performance measures, one measure of professionalism, and the two measures of satisficing behavior. Four segments are used corresponding to: (1) no error (Ideal), (2) one error (Typical), (3) two errors (Imperfect), and (4) three or more errors (Worst). In this context, it is the Worst segment which is of the greatest concern since it represents those who are most likely to give erroneous responses.

Results

Data is presented below for each test panel in clusters of three figures each representing a single panel tested for consistency. Each cluster of three consists of a “radar chart” that indicates which of the variables tested, when compared to the average of all waves proved to fall outside of two standard errors. For example, the panel in figure 1 indicates stability in 12 out of 15 variables. Three variables, all measures of hyperactivity among respondents: number of surveys in the past month, a survey every day and participating in 5 or more panels, appear to have changed over the course of testing. Critical variables like the three segmentations remained stable. Examination of these variables proved that they have been on the increase in this panel.

The data collected in Figures 1-3 is “blinded”; the panel was unaware of our analysis and gave us unfettered access to examine any respondents that we chose from a data base of pre-

profiled respondents which we had collected from this panel during a period of over a year. We consider this our most rigorous test. Similar data collected from this panel that was unblinded revealed identical results.

The consistency that we see in this panel is particularly acute. As an example, Figure 1 also contains a sequence of stack charts representing Buyer Behavior segmentation data, which in this case is extremely stable.

The third component of the Figure 1 cluster is a sequence of stack charts containing data on Q-Metrics®, a proprietary engagement measure explained above. We draw your attention to the top segment representing the “worst” respondents. In this case, these respondents represent less than 5%, a threshold typical of many online American Panels.

The data captured in Figures 4-6 cluster is in the same format as figure 1 but is compiled from a different American panel collected in a similar blinded fashion.

Figure 1 displays panel consistency for a blinded American panel, which by most measures would be considered to be very consistent. The “radar” chart indicates which variables fell outside of the acceptable bounds (Figure 1). The behavioral segmentations were remarkably stable, as evidenced by the buying behavior (Figure 2). In this case, the company was making an aggressive effort to reduce the hyperactivity of respondents as measured by the fraction of speeding, belonging to five or more panels, and participating in 30 or more surveys. The increase in “ideal” respondents over time (Figure 3) which includes, as one of its measures,

number of surveys per month, correlates with the inconsistency depicted in the radar chart.

When we measure consistency we note any change, even one normally thought to be beneficial.

Figure 1. Radar chart for consistency testing on a blinded American panel.

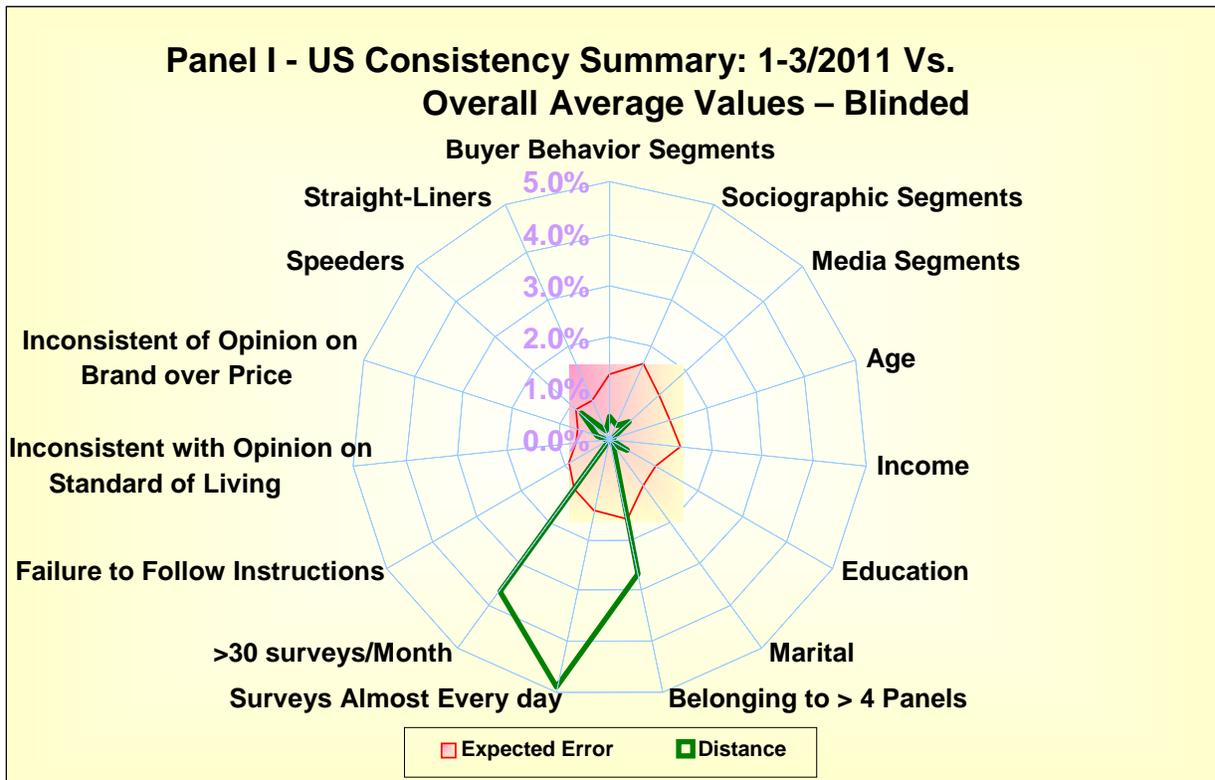


Figure 2. Stable buyer behavior segmentations for a consistency test on a blinded American panel.

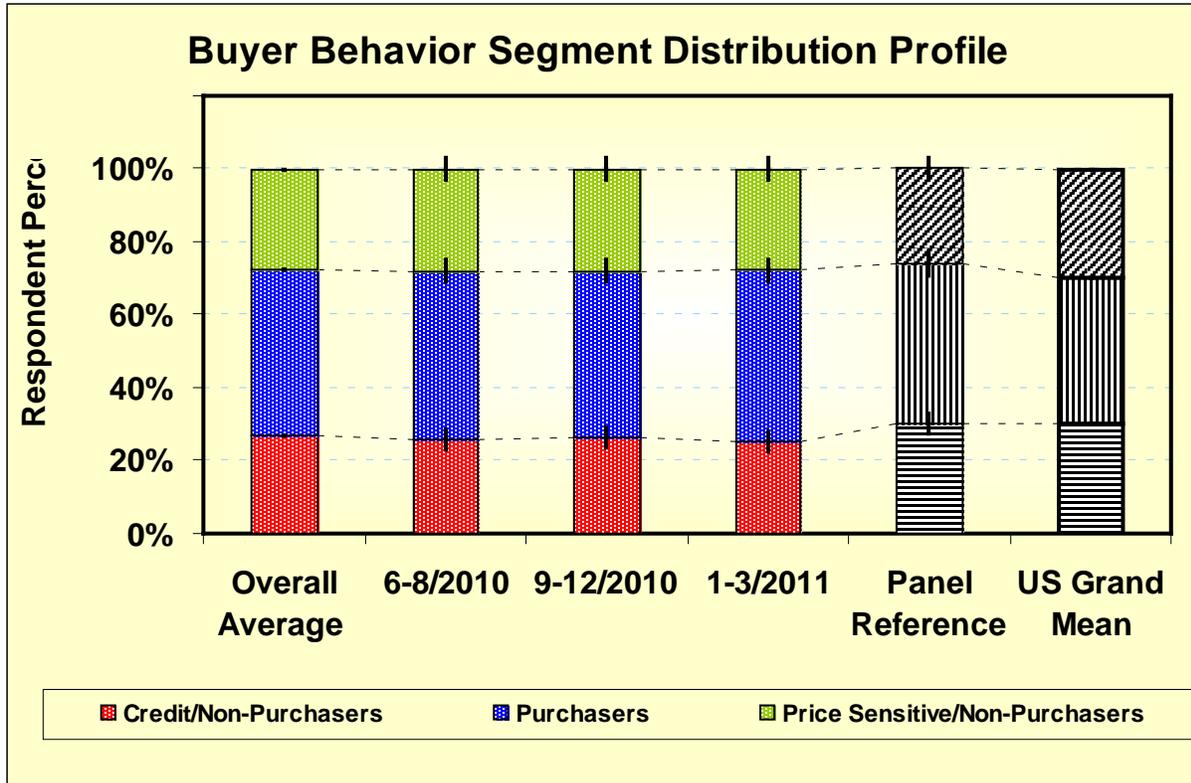
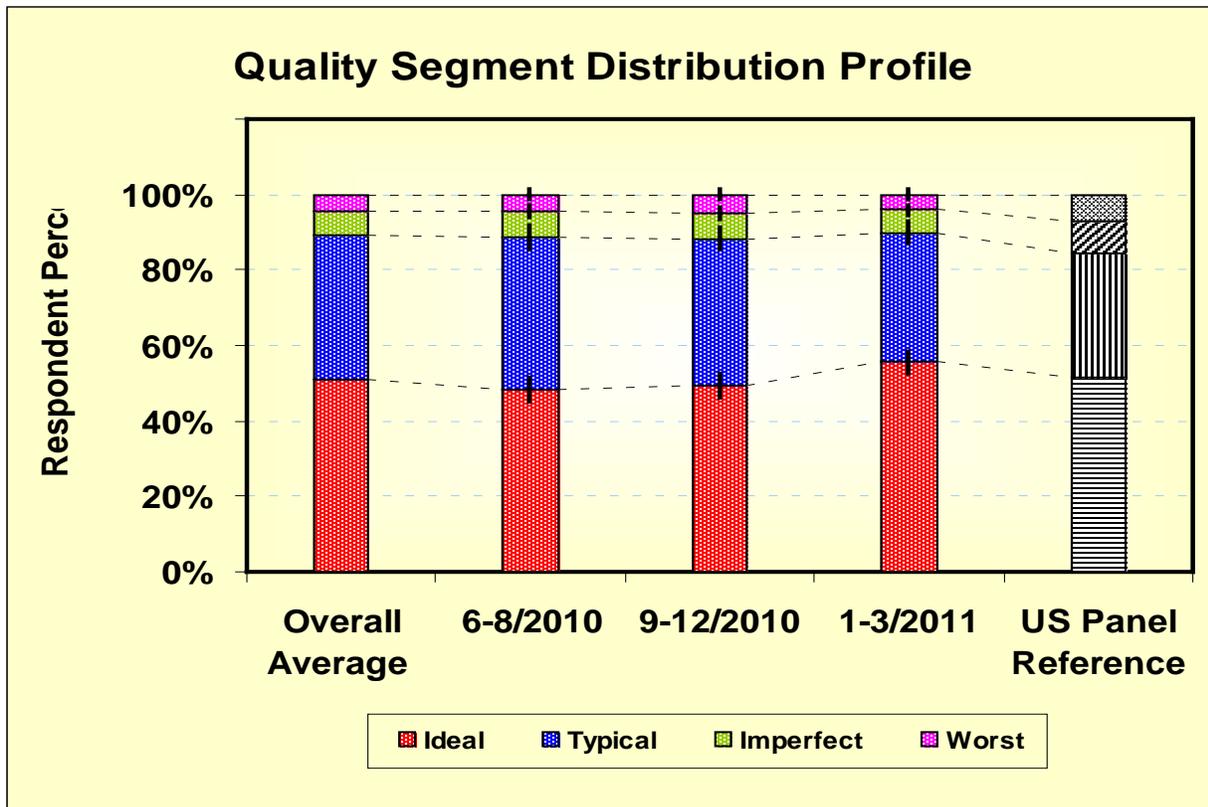


Figure 3. Quality segment distribution for a three wave consistency test by a blinded American panel. The top segment, which represents those respondents who failed three or more of six tests are considered the “worst” respondents and are often targeted for removal as their data is suspect. In American panels, the percentage of “worst” respondents has averaged 5%.



Figures 4-6 contain Panel Consistency for a second American Panel. Sample collected here is blinded. Notice that there is increased variability in the three segmentations and failure to follow the trap question. Here we would score this panel an 11 out of 15 measures and consider it stable as both education and taking a survey almost every day were both borderline. Further, as the number of waves included in the consistency analysis increases, the bar rises due to increased sample size. Notice that the QMetric “worst” segment hovers around 5% and that the buying behavior segments are somewhat more variable than those in Panel I. The increase in the two lowest-quality QMetrics segments and the increased variability of the buying behavior, suggests that there is a relationship between respondent engagement and the ability of a panel to be consistent.

Figure 4. Radar chart for a second American panel subjected to a blinded consistency analysis. Results are excellent and the data appears quite consistent as the RMS error hovers below 2%.

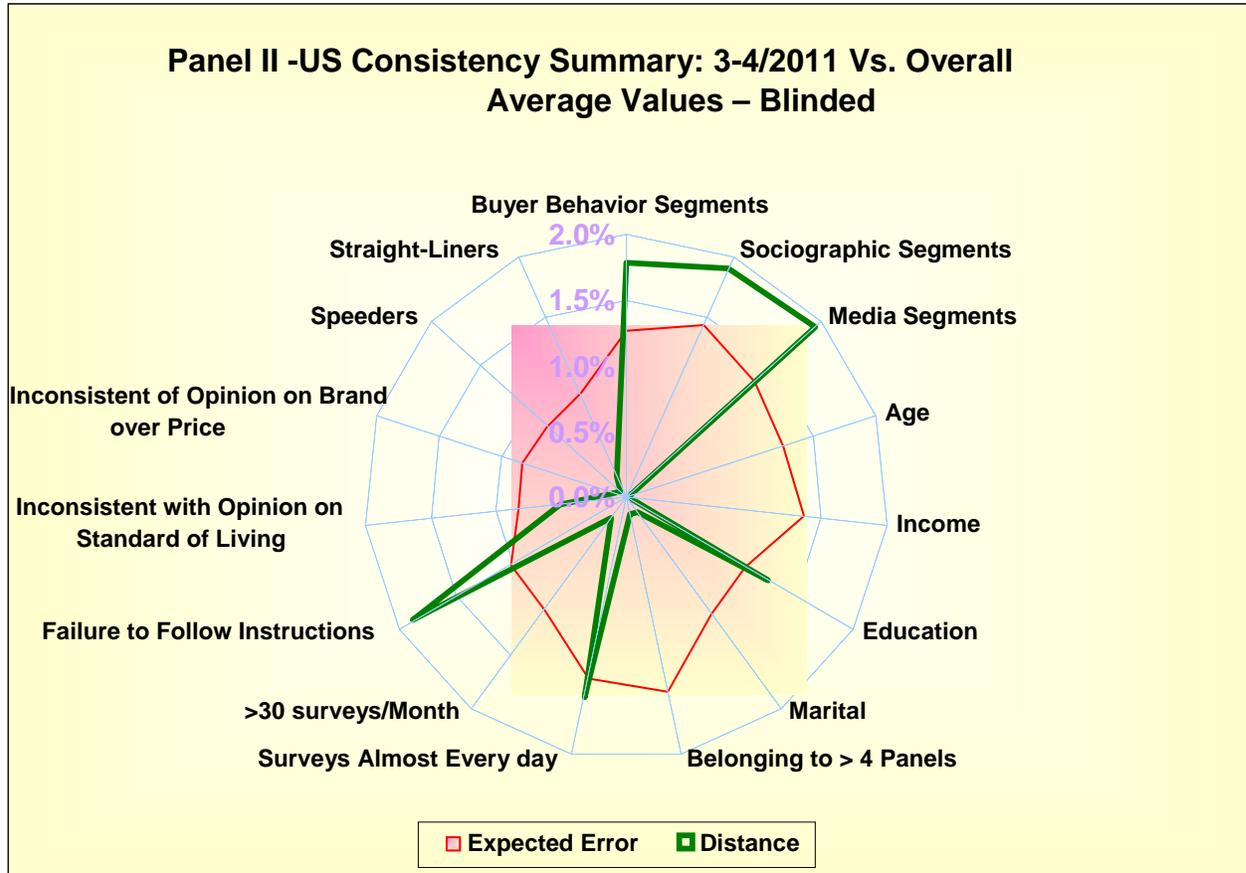


Figure 5. Buying behavior from a second American Panel subjected to a blinded five way consistency analysis. The Buying behavior is variable but not to the point that we believe it threatens the interpretation of data collected from this source.

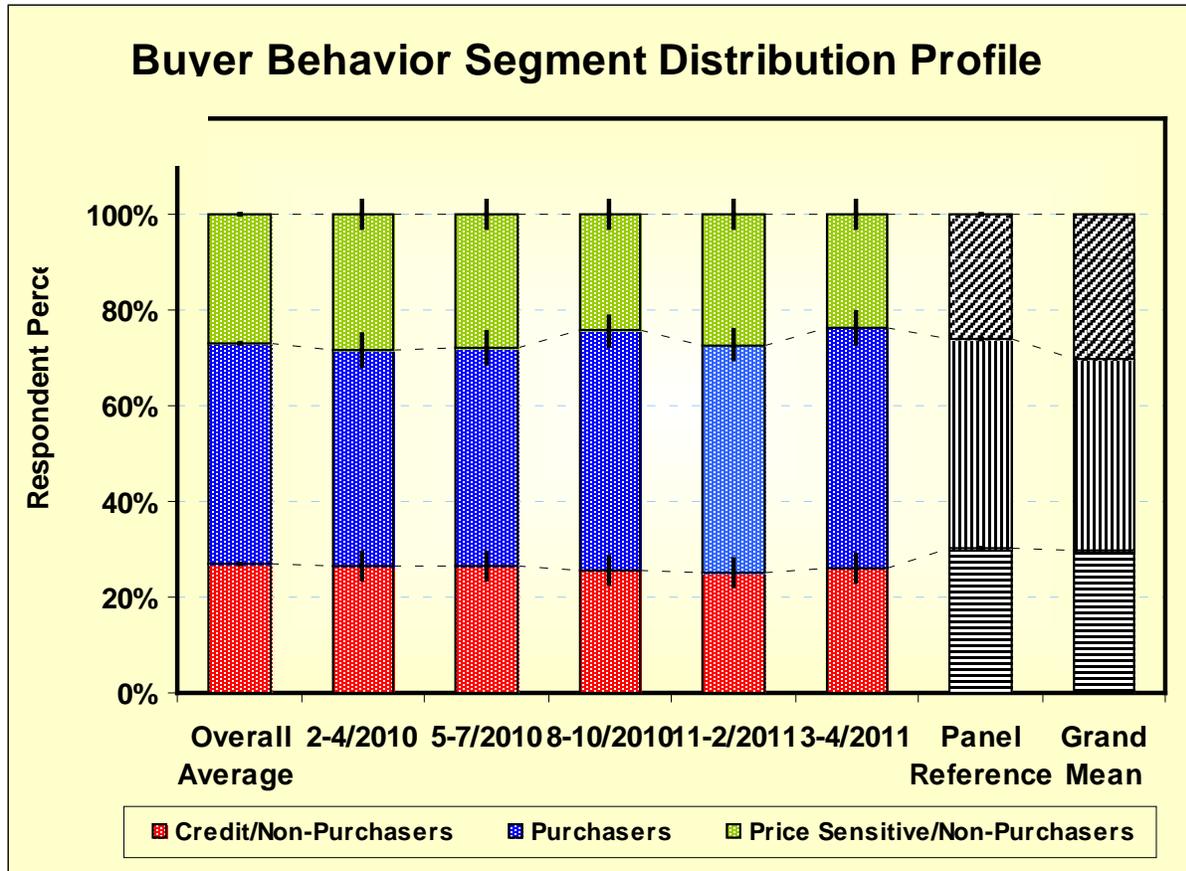
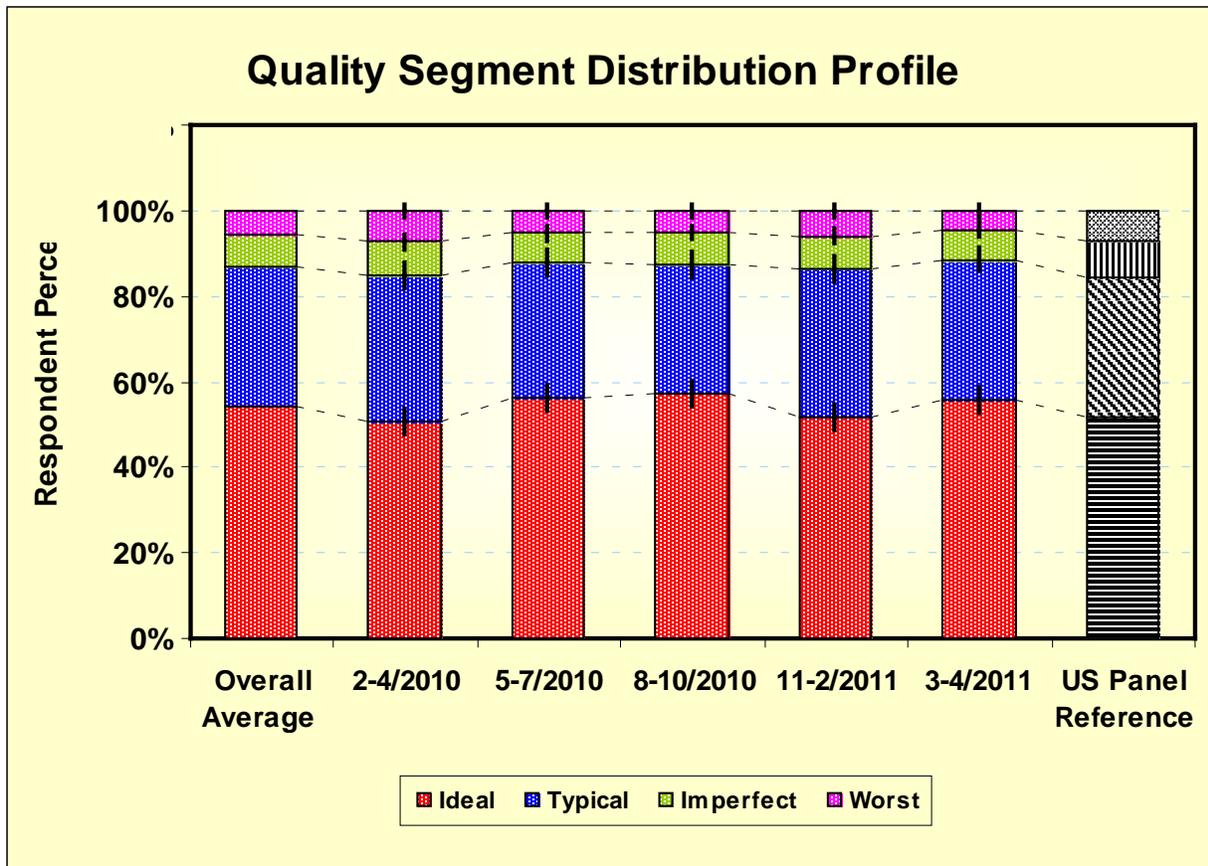


Figure 6. As might be expected from a panel that demonstrates consistency, the number of “worst” respondents is relatively low and consistent. This indicates little shift in the quality of panel respondents and perhaps a commensurate stability in the data. However, we note that the top two QMetric bars are greater than the previous panel.



Continued consistency analysis can diagnose subtle changes in panel management. Here (Figures 7-9) a usually consistent panel exhibited a shift in buying behavior. Discussion with management revealed that a sequence of subtle changes in the recruiting “pop ups” that they were using appeared to drive a buying behavior inconsistency. The vigilance of the panel managers helped us diagnose the very change in sourcing and correct the problem by reverting to the earlier means of recruitment.

Figure 7: Blinded consistency analysis in an American Panel. The inconsistency was diagnosed and corrected by modifying the respondent recruitment strategy.

Panel III -US Consistency Summary: 11/2010 Vs. Overall Average Values – Blinded

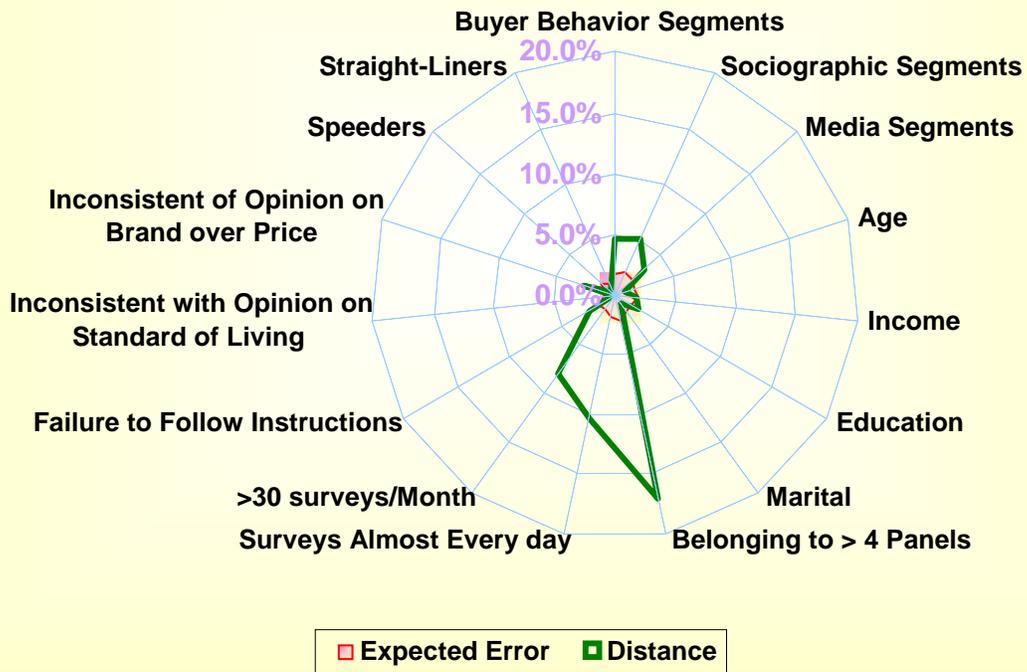


Figure 8: Buying behavior shifts due to recruitment banner wording changes.

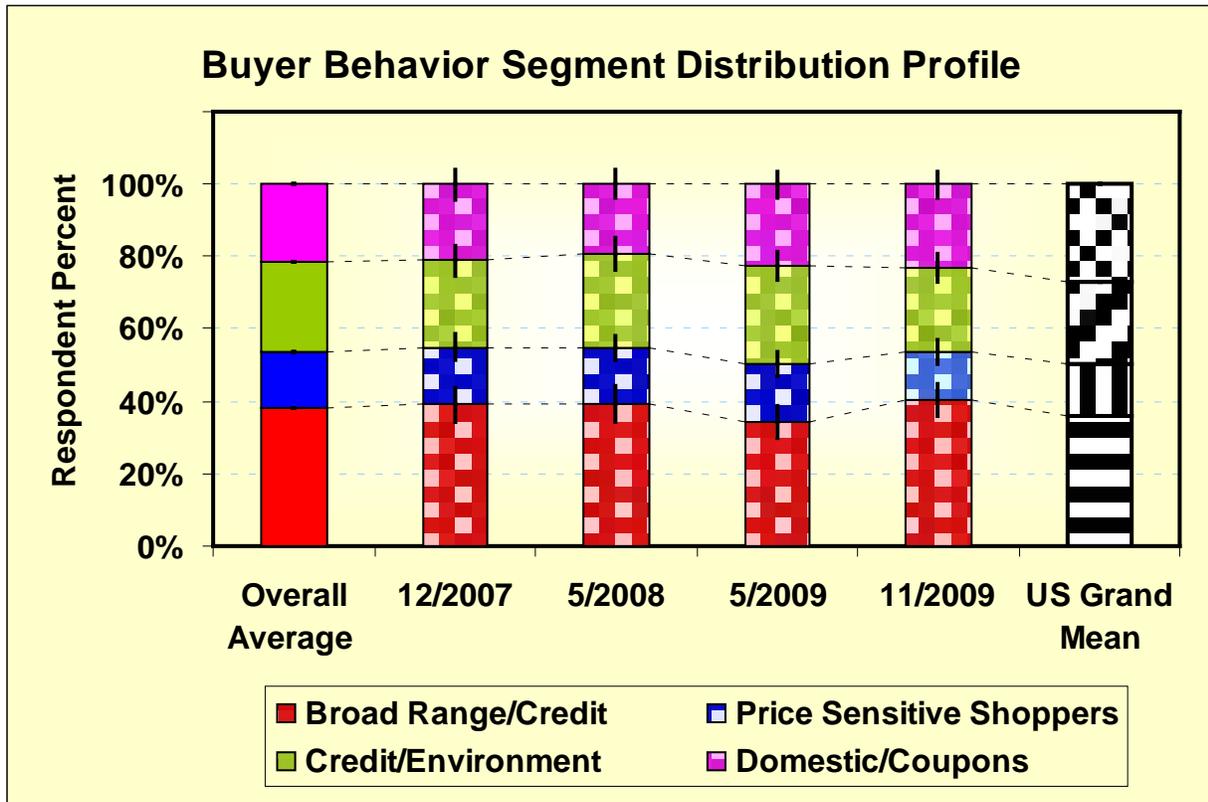
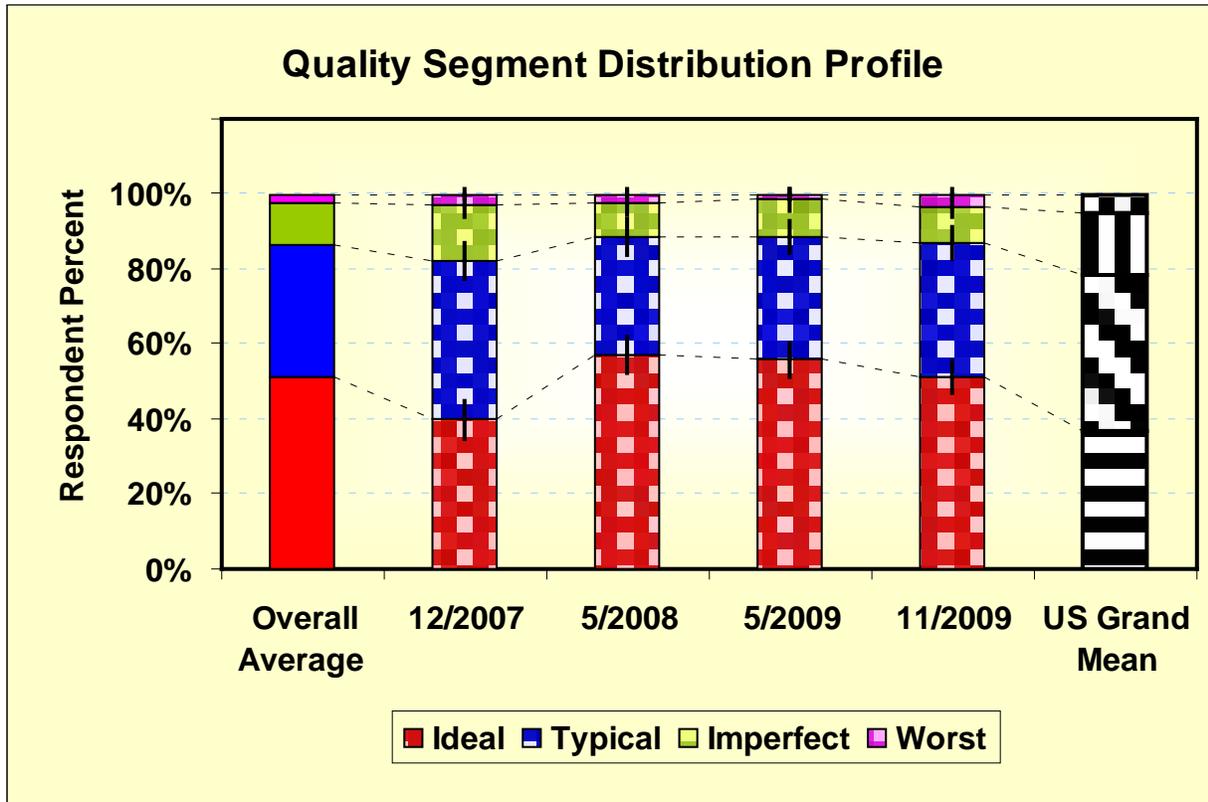


Figure 9: As the respondent engagement appears high and consistent, based on the top segment of “worst” respondents we had to look elsewhere to find a solution for the inconsistency evident in the 5/2009 wave.



One would think that panels would have a decided advantage when conducting an unblinded consistency test. If this were a tracking study the research data would be meaningless. In this case the inconsistency here is likely the outcome of changing sources. Panels that fail to understand the impact of sourcing are often the victim of their own largess (Figures 10-12).

Figure 10. Inconsistent on all three behavioral segmentations and struggling to meet quotas on age, this panel would be a poor choice for a tracking study.

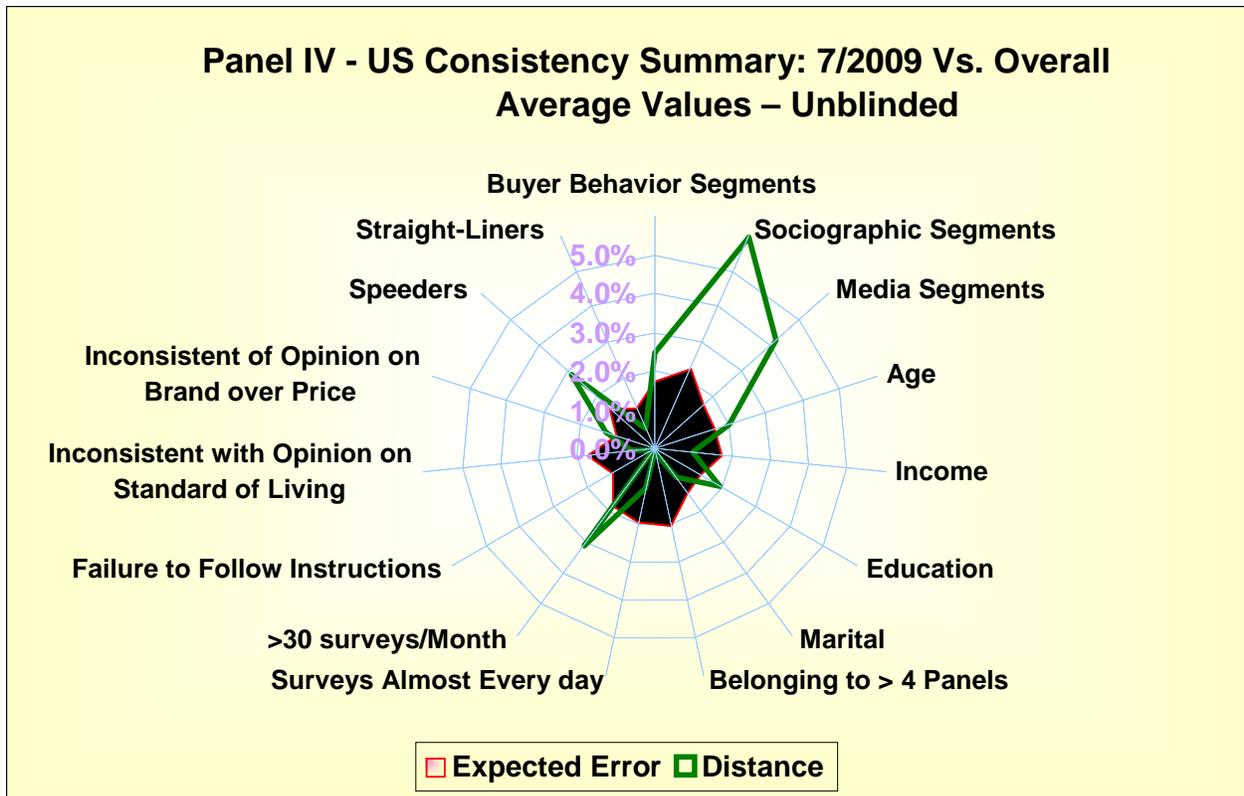


Figure 11. While the buying behavior was the least inconsistent of the three behavioral segmentations analyzed in this panel, clearly the changes exhibited between wave one and two are enough to disrupt data in any tracking study. Even a one-time study would suffer data anomalies based on the variability evident here.

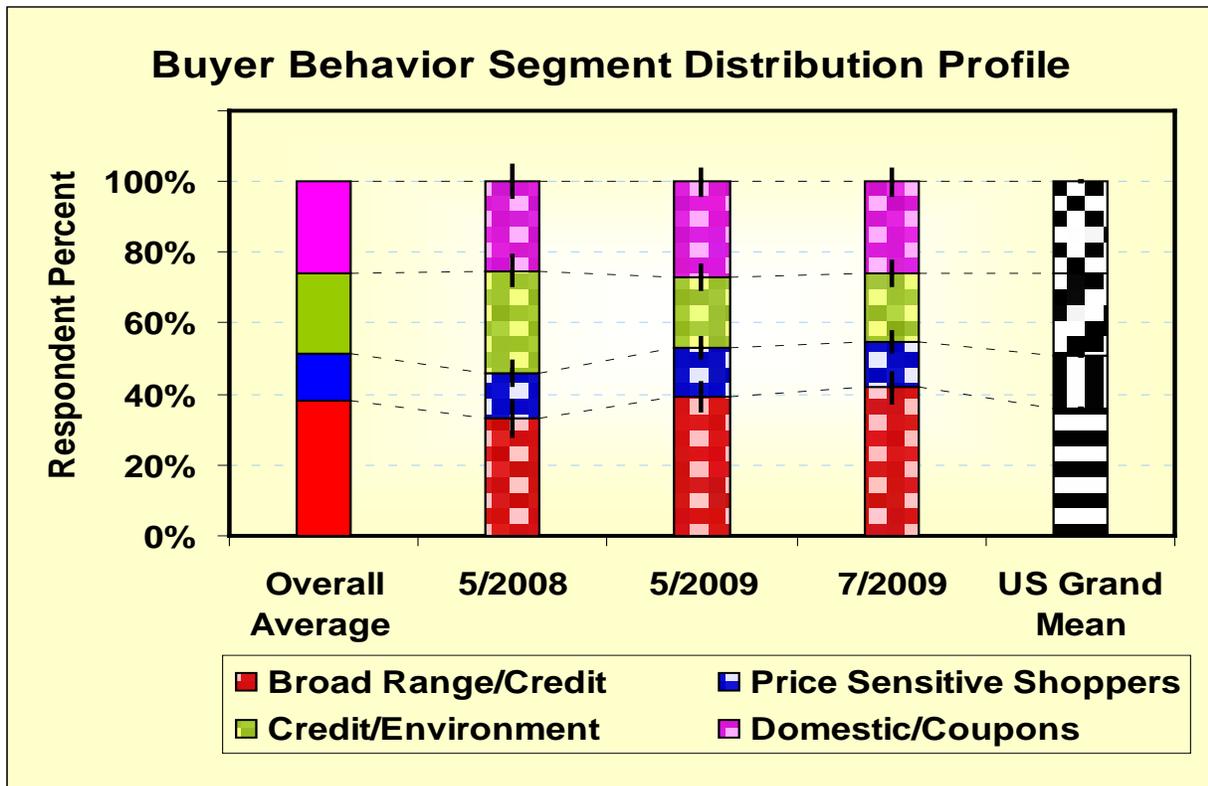
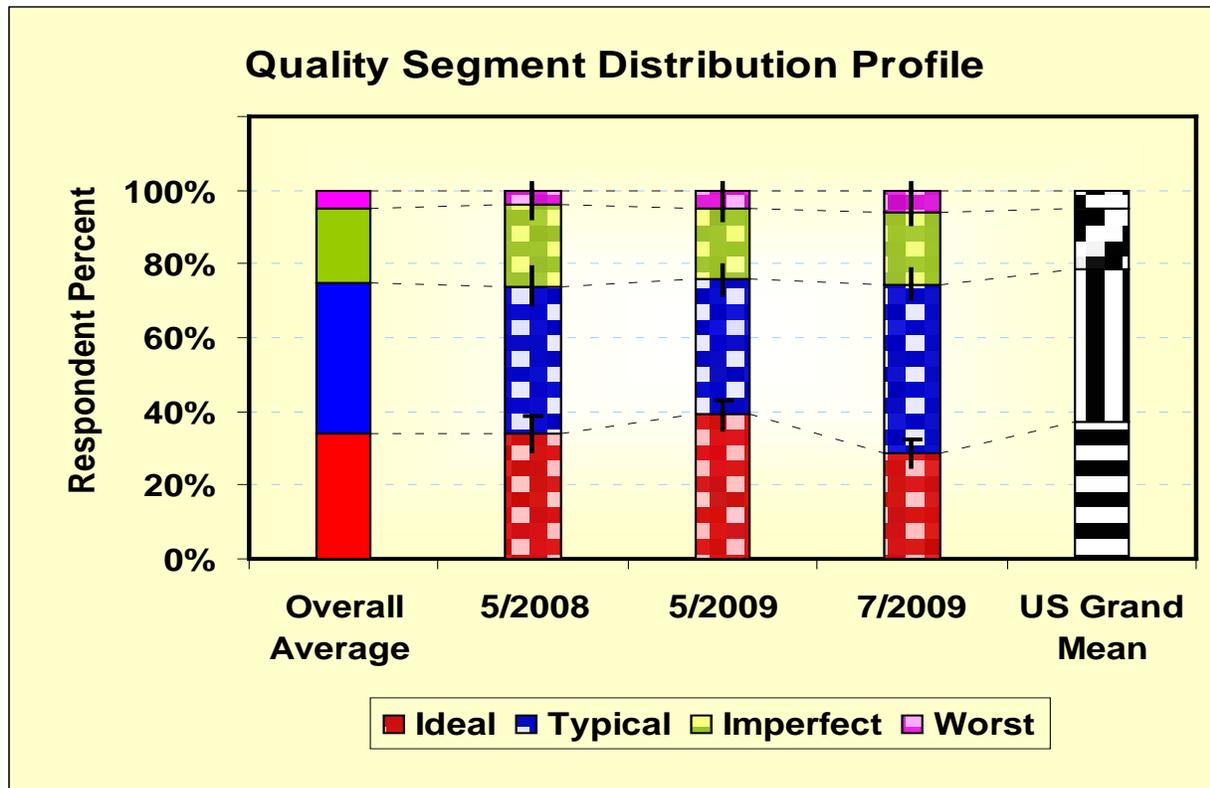


Figure 12: With the two lowest-quality metric segments in excess of 20% we find the inconsistencies reported previously to be unsurprising.



Stable sourcing is a driver of consistency. Here the sourcing is rather consistent and the respondent base is highly stable and long term Figure 13-15. Of interest is that the “purchasers” segment is somewhat low for an American Panel, as it usually hovers in the 44% range. As this panel is part of a particular buying community, this difference is not surprising. The differences, we note in the segment distributions of the panel, provide us with important tools for blending panels together. This particular configuration is rather unusual and closer to what we believe to be typical in the general (not just online) population.

Figure 13: A relatively consistent online panel. Here the consistency appears driven by consistent sourcing.

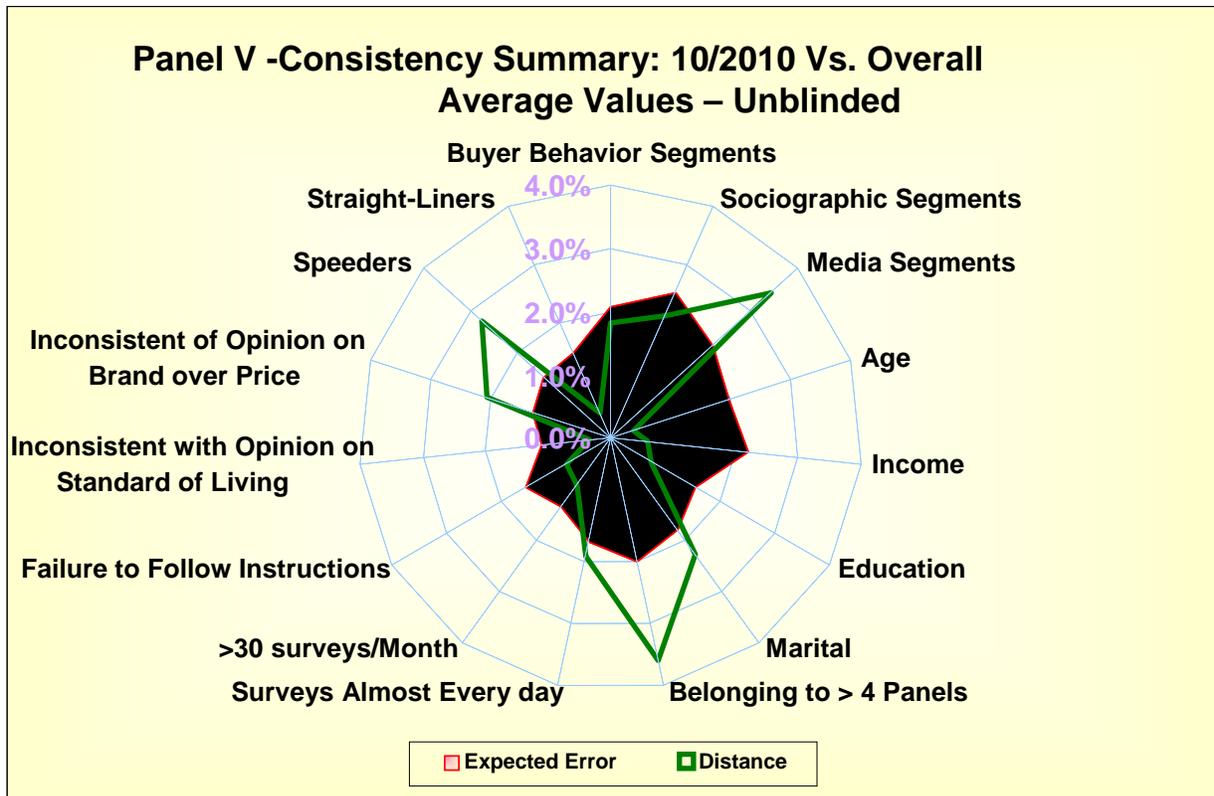


Figure 14: Buyer behavior consistency in a panel drawn from a particular purchasing community.

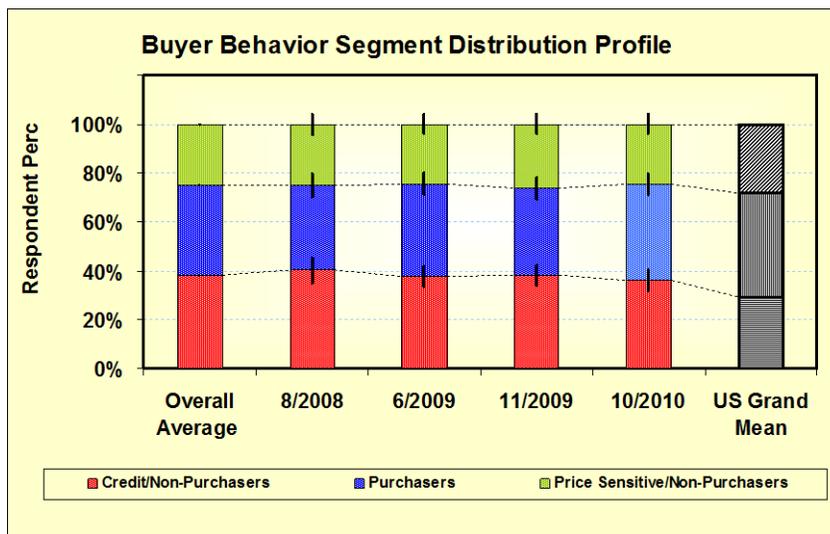
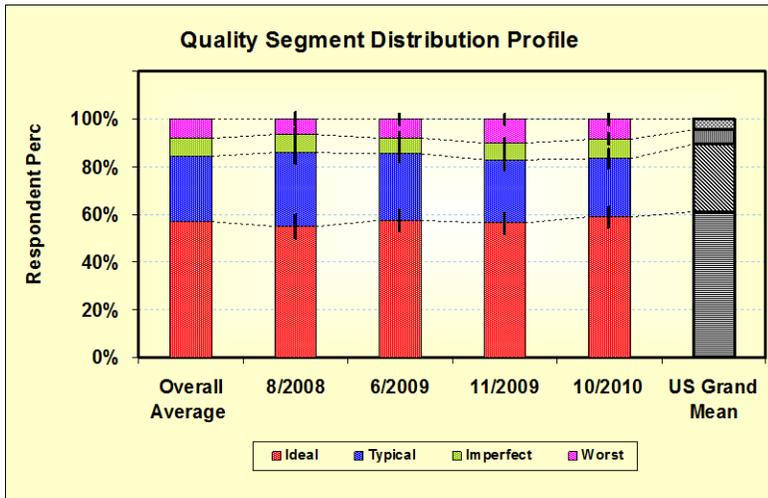


Figure 15: The two lowest-quality Q-metrics are approaching an area of concern. However, respondent engagement does not appear to have driven long term inconsistency due to the stability of sourcing.



Now let's take a look at consistency in a Canadian panel. Here the relatively low levels of unengaged respondents, coupled with stable sourcing generate a picture of consistency (Figures 16-18).

Figure 16. A radar chart of a consistent Canadian panel.

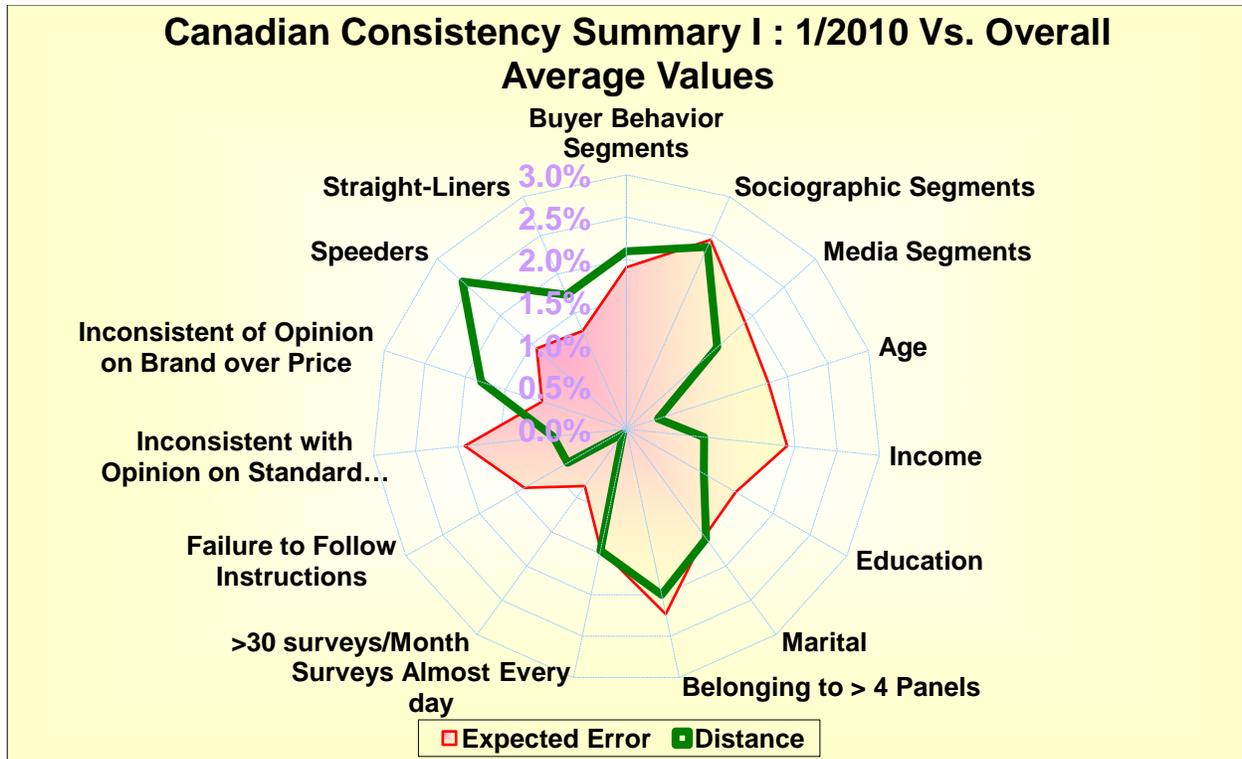


Figure 17: Relatively consistent buying behavior in a Canadian panel.

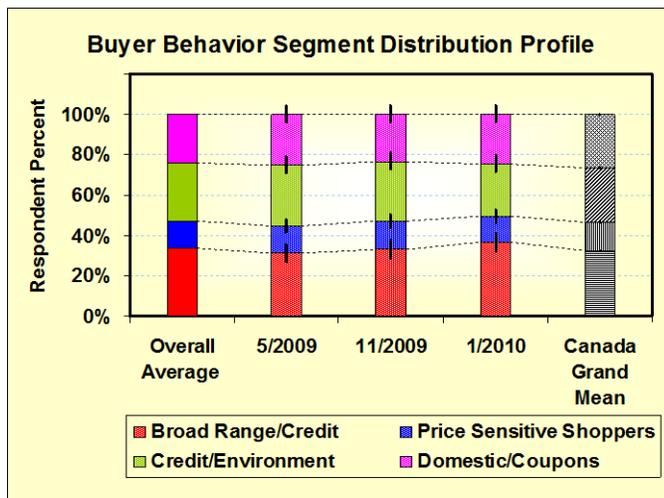
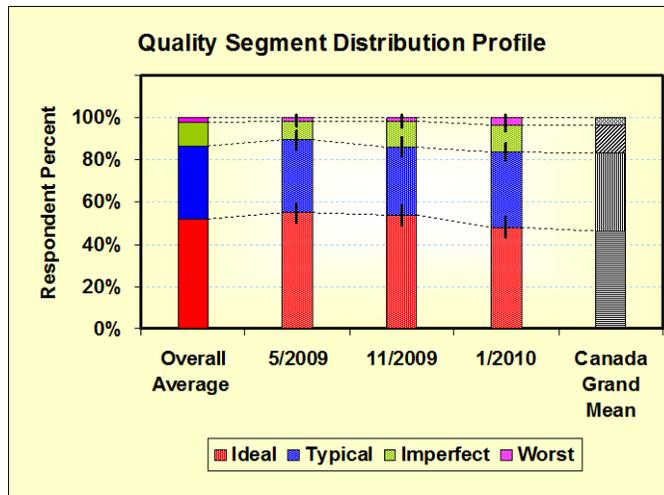


Figure 18: The two lowest-quality segments of QMetrics are increasing. However, the very worst respondents are very stable and extraordinarily few. It is unlikely that low levels of

respondent engagement will drive inconsistencies here unless the increasing trend of worst two segment infractors continue to increase.



The “BRIC” countries (Brazil, Russia, India and China) represent a significant challenge to market research data collection firms. Although this Brazilian panel (Figures 19-21) appears to have begun the analysis with highly engaged respondents, the two lowest-quality segments of respondents are rapidly growing. Concurrently there appears to be an inherent instability in the behavioral segmentations. We do not believe that this change is reflected in the population as a whole. These data indicate an increasingly unengaged population within the panel that is behaviorally in flux. This is likely to be due to a combination of increased respondent tenure coupled with the introduction of new sources. One would have to use extreme caution in launching studies with this panel.

Figure 19: Results of a Brazilian consistency analysis. Data on all three segmentations is highly unstable.

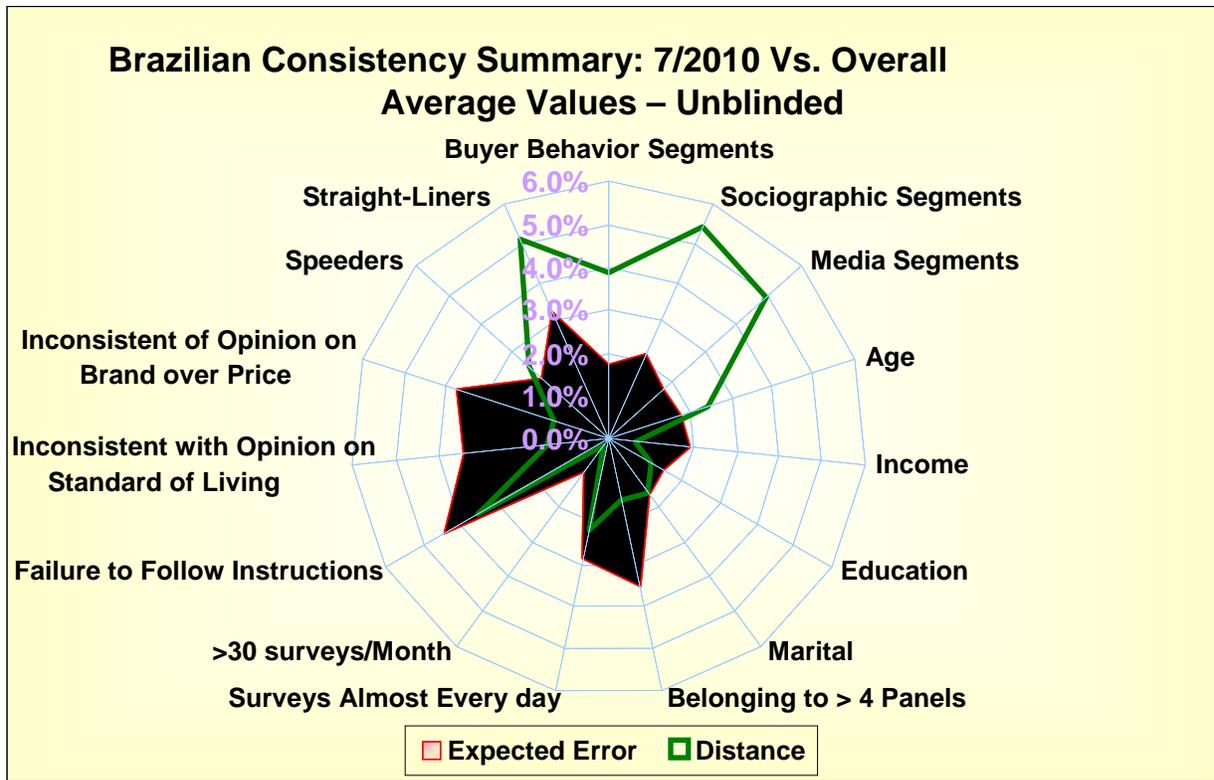


Figure 20: Shifting buyer behavior segmentations in a Brazilian panel.

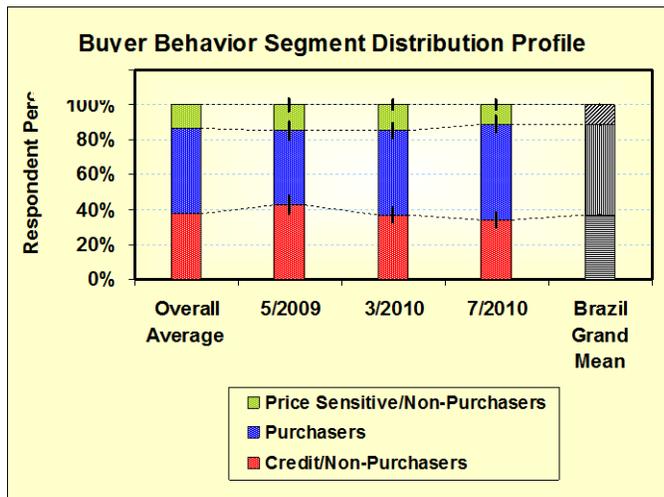
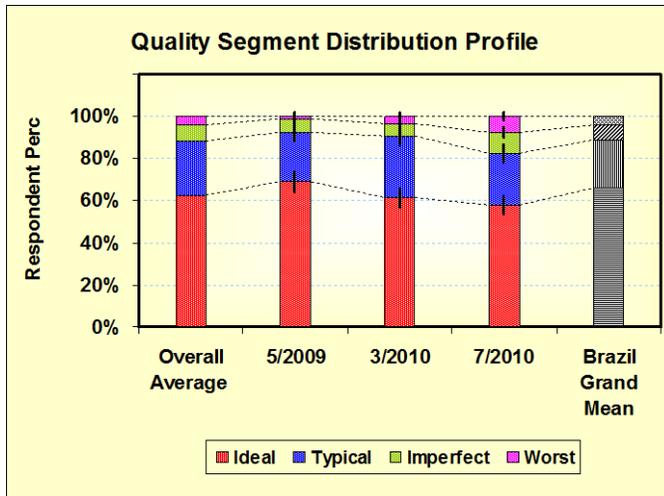


Figure 21. Increases in less engaged respondents appears to be the likely driver of panel instability in this Brazilian panel.



We find Russian panels to consist of highly engaged respondents (Figure 22-24). As the number of waves in a consistency analysis increases, it becomes harder to stay within the decreasing error bound. On the standards of analysis with fewer waves of testing, this panel is relatively stable. We note that buying behavior is variable, but of the three segmentations, it is the most problematic.

Figure 22: Consistency in a Russian panel.

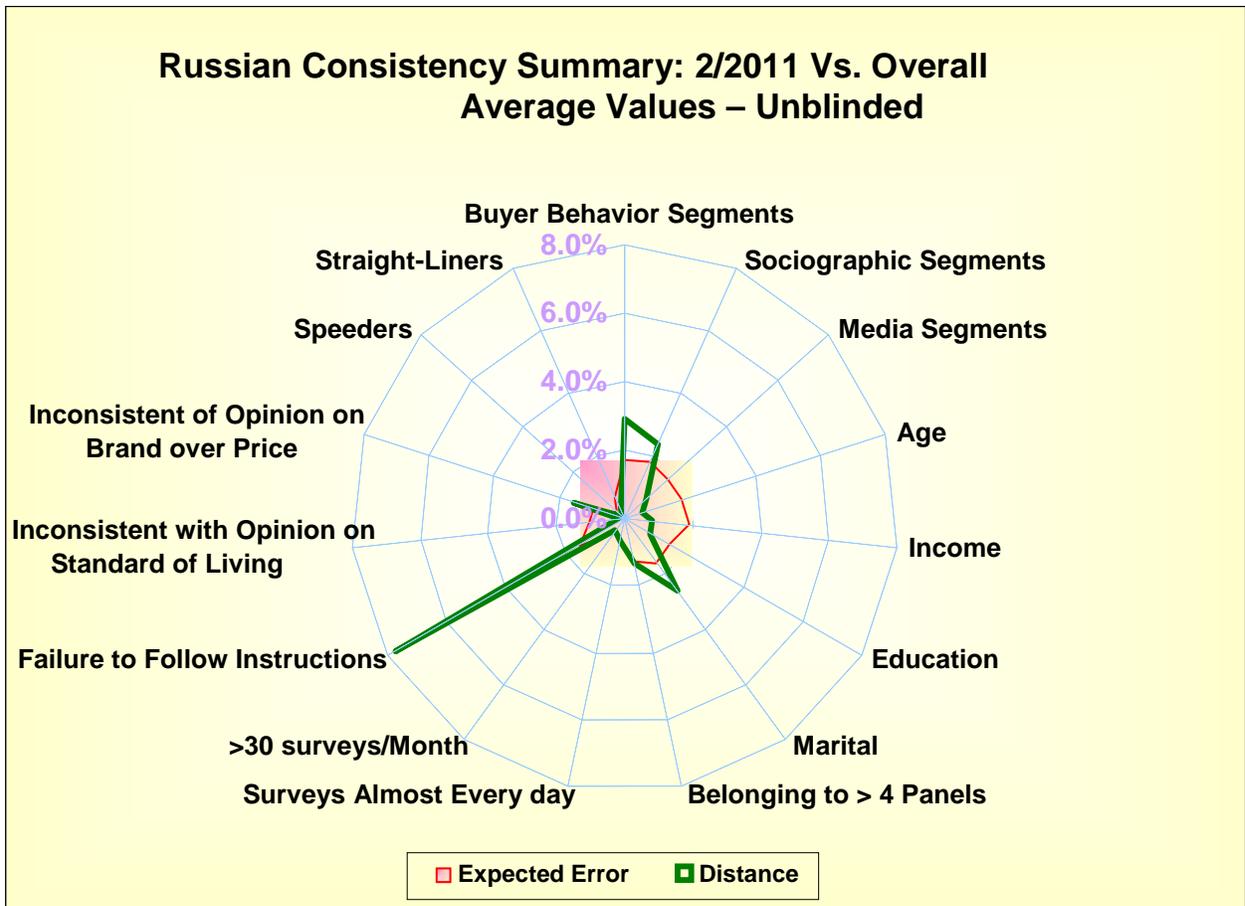


Figure 23: Consistency in a Russian panel.

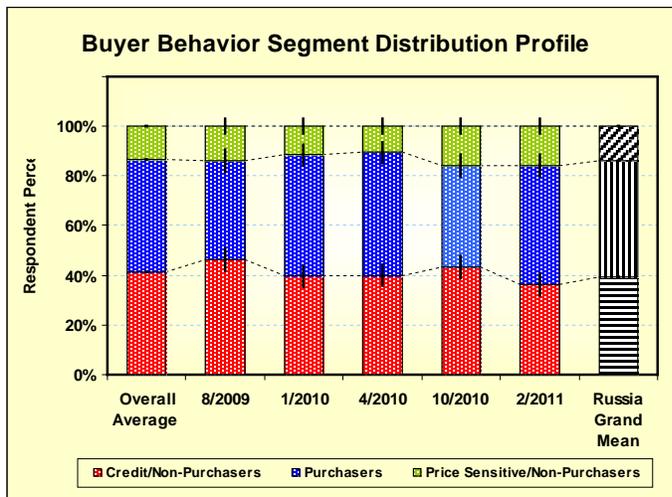
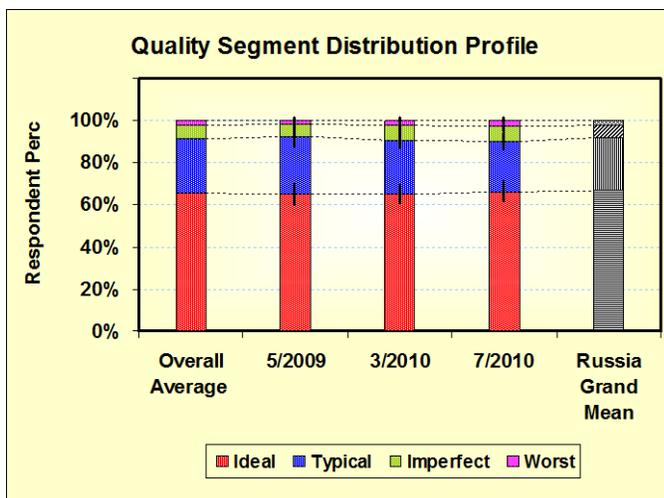


Figure 24: High levels of respondent engagement within Russian panel.



No country thus far confronts us with greater inconsistency than does in India (Figures 25-27). In this case, the participating panel made every effort to stabilize the data from wave to wave. But it's quite evident consistency was seemingly unobtainable. We doubt this was the product of management inattentiveness as seemingly every piece of sample was scrutinized.

We draw your attention to the QMetric scores, where we experience the highest percentage of any sample we have seen to date; increasing beyond 20% for the "worst" category. This indicates that the respondents are unengaged and are likely to satisfice. Under these circumstances, the data flow is unlikely to be stable and the panel source should have considerable difficulty in passing the hurdle of a consistency test.

If data instability approaches an almost random condition, the data obtained is useless if not dangerous for research practitioners to use.

Figure 25. Highly inconsistent results in an Indian panel. Panel management was aware that the test was being conducted.

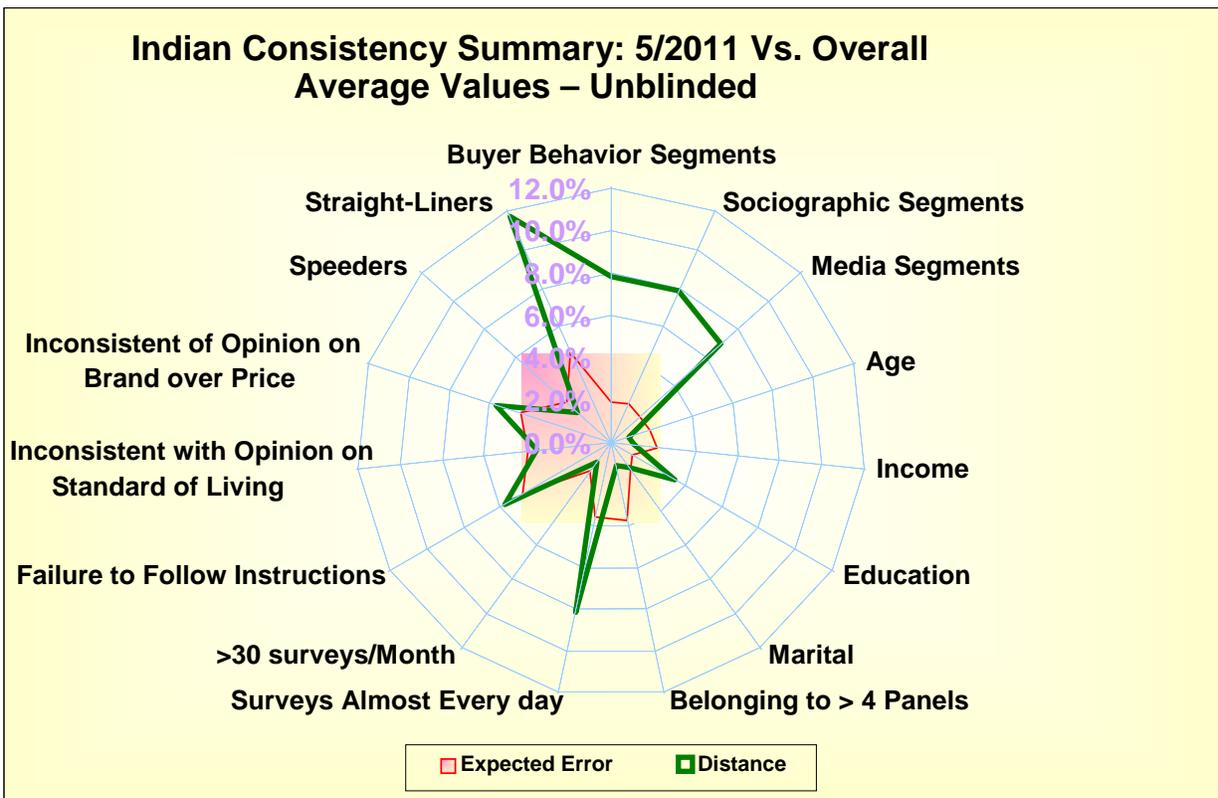


Figure 26. Inconsistent buyer behavior results in an Indian panel.

Consistency in Global Non-Probabilistic Online Samples

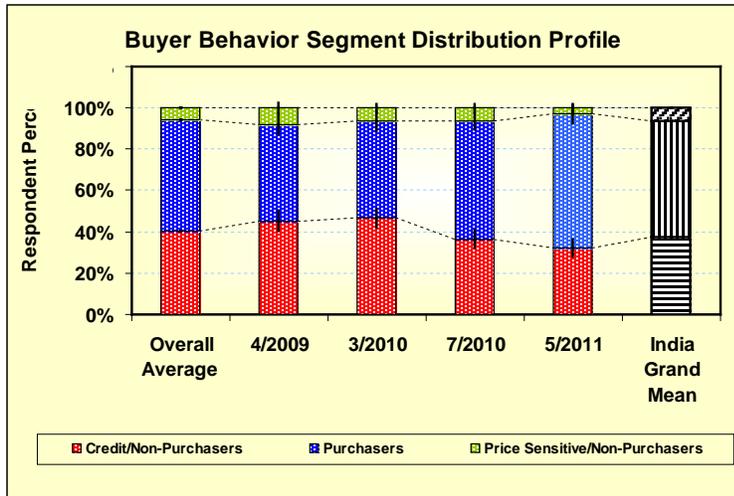
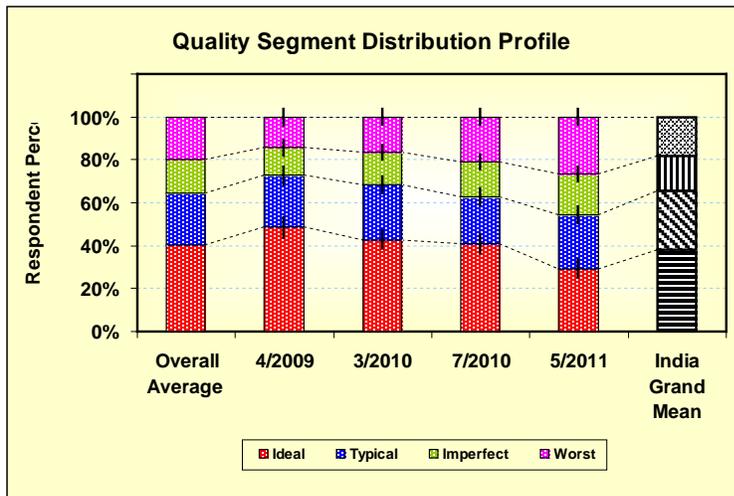


Figure 27. Extremely high level of unengaged respondents in an Indian panel.



Our hypothesis shows that when respondents are unengaged and enter a high frequency of random responses the sample source should have considerable difficulty in providing consistent results. In our Chinese consistency test, the participating panel was unable to meet the age,

Consistency in Global Non-Probabilistic Online Samples

income, gender quotas that we normally require (Figures 28-30). By restricting the demography, they in turn reduced the potential for variability. Thus their data appears consistent despite high QMetric scores. Given that the respondents show low levels of engagement and fail our battery of quality metrics to a high frequency, we question the value of the data collected and believe that the consistency demonstrated should not be accepted as valid.

Figure 28. An apparently consistent Chinese panel. As the demography was narrowly limited, we believe that the consistent results are an artificial artifact.

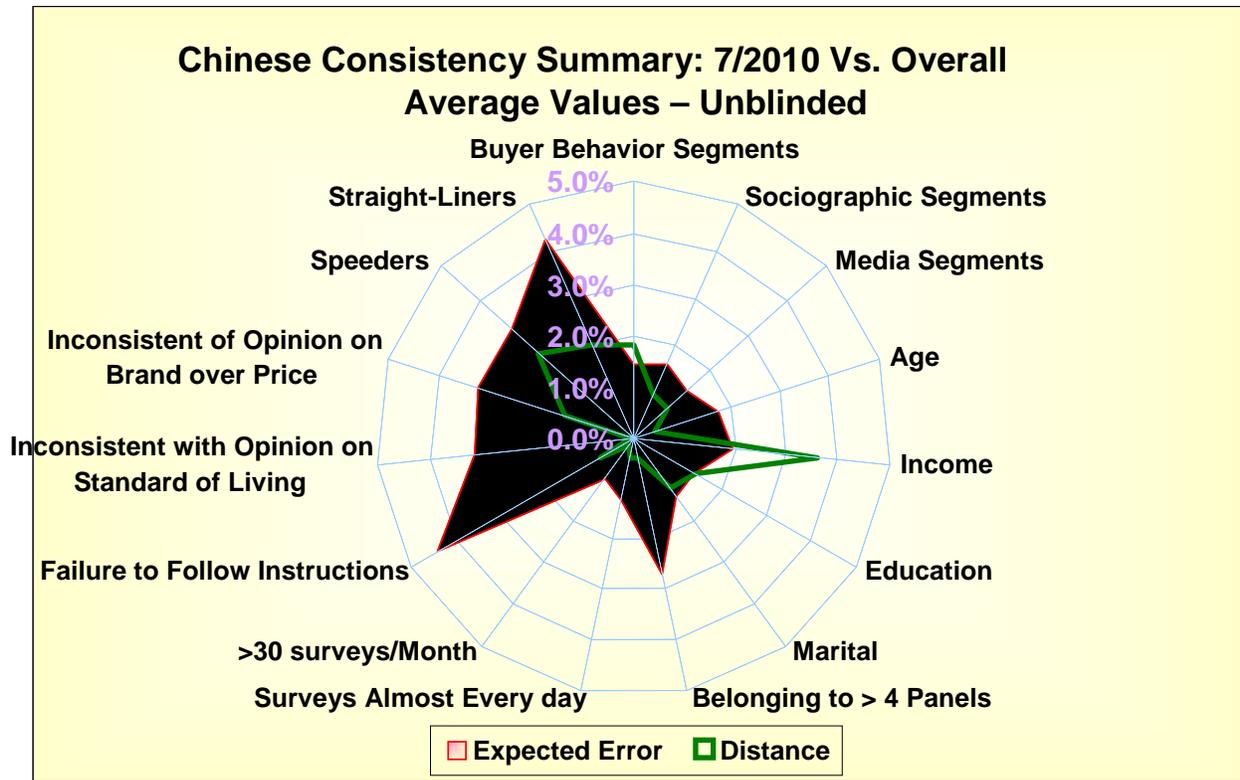


Figure 29. Buyer behavior in a “consistent” Chinese panel. Note that the price sensitive segment is almost non-existent. We believe this is due to a narrowly limited demography, allowed because the panel was unable to reach quotas.

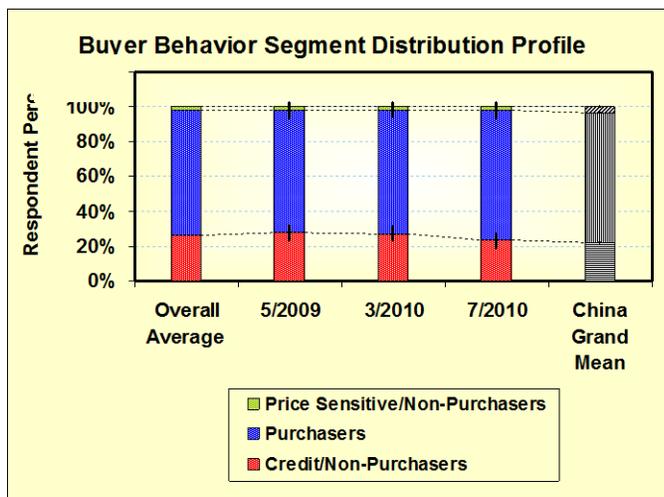
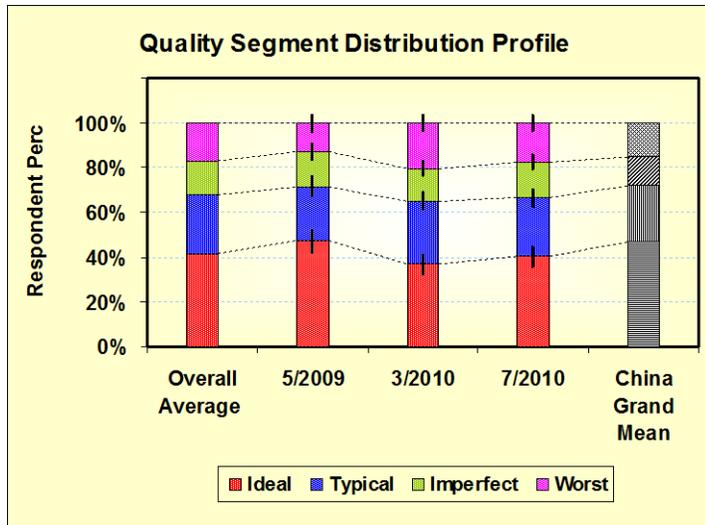


Figure 30. Extremely high levels of unengaged respondents in a Chinese online panel. This should have driven high levels of inconsistency as the data collected from such respondents is often random and poorly thought out.



Conclusions:

Respondent engagement is an important component of respondent quality. All respondents across these studies were administered identical translated questionnaires and yet levels of engagement vary drastically across panels. The causes of these differences are myriad and best explored elsewhere (as in Gittelman & Trimarchi, forthcoming) but perhaps the consequences are in question.

The importance of a consistent sample frame is easily grasped – a panel company must be able to provide consistent results to questions with stable population values or else changes in more dynamic variables will be difficult to interpret. In the pursuit of consistency, disengagement plays a vital role. If respondents are answering in any way randomly, it is extremely unlikely

their responses will be useful or replicable. Since this error is theoretically random and not systematic, it may not necessarily cause changes in overall means, but this error can manifest itself in fluctuations that could easily be misinterpreted as meaningful. In essence, researchers must know if the changes they see in their data are real or artifacts of an inconsistent sample frame.

This relationship is not mere conjecture. Across the twelve panels tested for consistency, the incidence of low quality respondents (those marked for two faults or more in our QMetrics scale) correlated with variance over time with $R = .5$, which rises to $.75$ when discounting the Chinese market. While the Chinese data implies consistent behavior despite very low respondent engagement, it represented a very poor cross-section of its population. We conclude that the insufficient variance present in income, age, or education levels created a false homogeneity that is reflected in China's unusual behavioral profile and likely contributed to above-average though misleading consistency.

References

Gailey, Ron. 2008. "The Quest for Sample Integrity: Exploring Panel Sample Characteristics That Impact Survey Results," Presentation at The Research Industry Summit: Solutions that Deliver Quality, IIR, Chicago, November 2008.

Gittelman, Steven and Elaine Trimarchi. 2009. "On the Road to Clarity: Differences in Sample Sources." Casro Panel Conference. New Orleans. 2009

Walker, Robert, Raymond Pettit, and Joel Rubinson. 2009. A Special Report From The Advertising Research Foundation: The Foundations of Quality Initiative, A Five-Part Immersion into the Quality of Online Research." *Journal of Advertising Research* 49: 464-485.