

Metrics for panel combination: a non probabilistic platform

Steven H. Gittelman

Survey research has historically relied on a probabilistic model to underlie its sampling frame. With rare exception online research is non-probabilistic. Research without the safety net of a probabilistic frame raises all kinds of alarms. Challenges as to the reliability of online research has become a growing crescendo as its non-probabilistic nature has become evident. However, not all sampling frames must or can be probabilistic. To insure that non-probabilistic sampling frames exhibit appropriate levels of continuity, predictability and reliability, they must be measured over time. Unfortunately, no such standard metrics exist to track reliability in online sampling. In fact, whether they are access panels or social networks there are no standardized means of balancing or comparing panels, until now.

We are proposing two practices: (1) using respondent behavioral metrics [professionals, speeders, straight-liners, inconsistencies, etc.] for which we have developed standard quantitative measures, and (2) the use of structural segmentations (media, purchasing and psychographics) of key variables as standards. We emphasize the need for two key requirements in our standard panel metrics including: (1) The ability to create a database that is prospective in that new sample sources can be added to the database without repeating the analysis and (2) a focus on indices that are pragmatic in their measure, for example we always view buying behavior as critical.

In this paper, we propose to use segmentation analysis as a new metric that will allow us to anchor online data in a new non-probabilistic sampling frame. Our goal is to use segmentation to create a fingerprint that can be consistently maintained by blending panels. By minimizing the variability of the segments through optimization and panel combinations, as compared to a standard, the sample can be stabilized. We cannot stabilize online data unless we provide it with a reference point to anchor itself; the segments help form that anchor. As the sourcing models continue to shift, panels will age and shift with them; we need a reliable anchor that rises above these problems. It is essential that we explore tools to measure these changes. Without a means of comparison, we cannot expect to measure drift nor can we expect to have a platform for predicting the future.

To define this reference point, we have launched a data collection effort in forty countries around the globe. In each country, we are collecting data from all willing respondent databases using a standard questionnaire similar to the one we used in the United States (Gittelman and Trimarchi, 2009). Each sample source provides 400 interviews with quotas on income, age, and gender appropriate for that country. We use native language translators to assist us in modifying our instrument for local use. The objective is to amass a minimum of five data sources in each country. We anticipate that the first wave of data collection will be completed in summer of 2009. Currently (April 28, 2009) we have completed some 36,000 interviews and have nick named the project the “*Grand Mean*.” In the United States we have collected data from 20 sources (8000 interviews) and have developed our modeling strategies on this data.

1. DATA SOURCE METRICS

To be effective, performance metrics must be standardized and independent. While there is always some degree of arbitrariness in the choice of metrics, their structure cannot be arbitrary. Metrics need to be standardized in such a way that analysis by any party would produce the same results. Furthermore, the procedures should be such that multiple executions will produce statistically common results. That is, that the error will always be within statistical limits.

Independence

Independence of the metrics is a key principle in their development. It makes little sense to have standards that are, in fact, tied in principle to specific types of data, methods, and instruments. An alternative way of looking at the issue of independence is that of objectivity. That is, that all usable standardized metrics must be objectively obtained. We believe that requires independence in several ways.

- ***Sources*** – Metrics must be independent of the panels and sources. They have to be applicable across all or at least most sources of data. Furthermore, the metrics cannot depend on the nature of the sources or in their construction.
- ***Methods*** – The methods for constructing the metrics must be independent. We are looking for different measures; performance measures should not be structurally interconnected.
- ***Instruments*** – The fundamental metrics should be extendable to any number of compatible instruments (questionnaires).
- ***Evaluators*** - In order to assure objectivity and independence of the metrics, it is critical that the evaluator needs to also be independent. This is required both as a means to assure objectivity and as a means of certifying that independence.

Methods Used

The testing procedures were based on the execution of a standard questionnaire. The questions were selected to allow a consistent standard and independent assessment of the panel or data source. In all cases, samples were made by the source supplier using their standard methods of management and incentives to reflect current operations. Four hundred completed interviews were used per panel or data source. This was done using a standardized online (internet form) questionnaire. This questionnaire is available in multiple local languages for global execution.

Typically for “random” and non-structured samplings some form of quota corrections are used to help balance the resulting sample. These quotas may reflect the specific needs of a study, regional or preference requirements or to balance demographics. In the latter instance, the objective is to correct for inherent incongruities against some standard, often the general population. Quotas were imposed for age, income and gender, while suppliers were requested to use “best practices” to balance region and no instructions were given regarding employment, education or other demographics.

The reference or basis of comparison for all metrics for this test was the Grand Total. This is viewed as a *Grand Mean* value. We used commercially available online access panels. Alternative sources of data such as random phone dialing were not included in the formulation of this *Grand*

Mean. It is important to also note that this *Grand Mean* is specifically regional. That is, it reflects the samples for a specified country and is not global.

2. STRUCTURAL SEGMENT DISTRIBUTIONS

Typically, panels and lists are filtered to balance demographics against some external standard such as the known general population. However, this still does not assure that the source reflects the targeted group of respondents or even the larger population. Valued segmentation is designed to capture the distribution of alternative groups of customers that are expected to impact the studies executed using these sources of data. They reflect the potential bias or hopefully the lack of it in the panel. Ultimately, the goal would be to balance panels to better reflect the target population distribution of these segments.

Pragmatically Valued Metrics

The key issue is to identify driving factors that need to be addressed to assure an effective sample. These are “pragmatic” issues in that they represent concerns regarding the makeup of panels and sources as they would affect studies. Here we use the term “pragmatic” in that it is not a theoretical construct of what should happen, but what is expected to happen. Studies involving purchases will depend on the distribution of “buyer behavior” segments and it is for this reason that it is the primary focus in these analyses. However, there are other pragmatic approaches that are of value, for example media preferences could be the driver for a media company and sociographic analysis can be critical in public opinion data.

Quality Clusters

In order to construct a functional structural segmentation scheme, the underlying groups should be based on the identification of high quality clusters. Whether it is based on cluster analysis or any other method of identification the underlying structure should be:

- ***Unique*** - The cluster structure should be singular, and unambiguous.
- ***Distinct*** - Distinction is measured by the degree that the variable values of segments are different from the *Grand Mean* and that of other segments.
- ***Reliable***¹ – Reliability reflects the clarity of the assignment into clusters; that is there is little ambiguity in the assignments.
- ***Functional*** - The resulting clusters must represent an important and varying structure that is meaningful to the process of collecting intelligence. As such, the structure should indicate variation within countries for which it is expected to change. That is, the definition of the clusters should not be a trivial universal principle which does not reflect variation.

¹ Reliability is tested as part of the formation of the segmentation model using Multinomial Logit Regression. These Logit models are then used to assignment segments with datasets not used in the original segmentation exercise. The existence of a stable highly reliable regression model is critical for the use of segmentation as an analytical tool for progressive sets of data.

Identifying Segments

The process of identifying structural segments can be thought of as having four steps, going from the selection of the variables through identifying segments and developing a regression model. This task is done with a substantial set of data, within a single country in order to provide a stable structure. The resulting regression model is then used to assign segments for all other datasets. As such, the resulting regression model represents the segmentation scheme. It is therefore a critical focus of the analysis and testing.

Select Variables ⇒ Cluster Analysis ⇒ Logit Regression Model ⇒ Test Results

The requirements for an acceptable structural segmentation scheme are formidable in that the resulting scheme must be highly distinct and reliable. This demands that the resulting model provides a clear assignment of respondents to segments. While statistical cluster analysis methods are very robust and almost any variables set can be used to identify segments, not all will meet the stringent requirements needed for these applications. As such, this process can be iterative where groups of variables are tested until a satisfactory set are identified.

1. **Selecting Variables** - Traditional cluster analysis abhors missing data and as such questions that either lack data or contain “don’t know” responses are usually excluded. Metric variables are preferred.
2. **Adjusting the Data** - It is useful to have all variables monotonic² and balanced. That is, preferred values going in the same direction. Furthermore, it is also useful to transform categorical data into combined metric variables³. For use with traditional cluster analysis methods, the data is normalized across the respondents. Otherwise, the techniques tend to focus on total average values rather than strength of particular variables.
3. **Hierarchical Clustering⁴ of Normalized Sample** – This is used to help identify the number of clusters that would be needed⁵.
4. **K-Means⁶ Clustering of Normalized Data** - All the normalized data is then clustered based on the number of selected clusters. This generates segment assignments to each respondent record.
5. **Profile Segments (Non-normalized Data)** - Summary statistics for the total database and for each of the segments is then generated to determine the distinctness of the segments and determine a description of them. This is done based on the original

² Changing the direction of questions (best to worst) is often useful for testing consistency and can be required for specific methods of data collection.

³ For example, we combine identified purchased products into a single variable of number of such products purchased.

⁴ Hierarchical Clustering used Wards linkage with Euclidean distances.

⁵ The tools associated with this approach gives insight into the impact of the number of cluster, including cluster trees.

⁶ K-Means clustering is used here due to the large dataset.

altered data (not normalized). These profiles are used to assess the quality of the segmentation.

6. **Determine Multinomial Logit Model** – Logit (Logistic) Regression is a non-linear curve fitting technique used with a categorical dependent variable. In its full form, it generates a probability or likelihood of a respondent being in a specific group. It is used here to develop a “progress model” which will be used to assign respondents to segments for datasets not used for the original segmentation.
7. **Test the Regression Model** – Because of the planned projected use of the Logit segment model, it needs to be extremely accurate in assigning segments. This reflects the distinctiveness of the segmentation scheme and its reliability. For this use, we required almost a 100%⁷ recapture of the original assignments by the model.

Structural Segmentation Definitions

Three segmentation schemes are being used in this evaluation focusing on: (1) Buyer Behavior, (2) Sociographic Factors, and (3) Media Use Factors. These are not the only segmentation schemes that can be developed for this process. However, these were well supported by the test questionnaire and fulfilled the requirements as structural segments.

2.1. 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 are shown below in Figure 1 and reflect the response to 37 input variables.

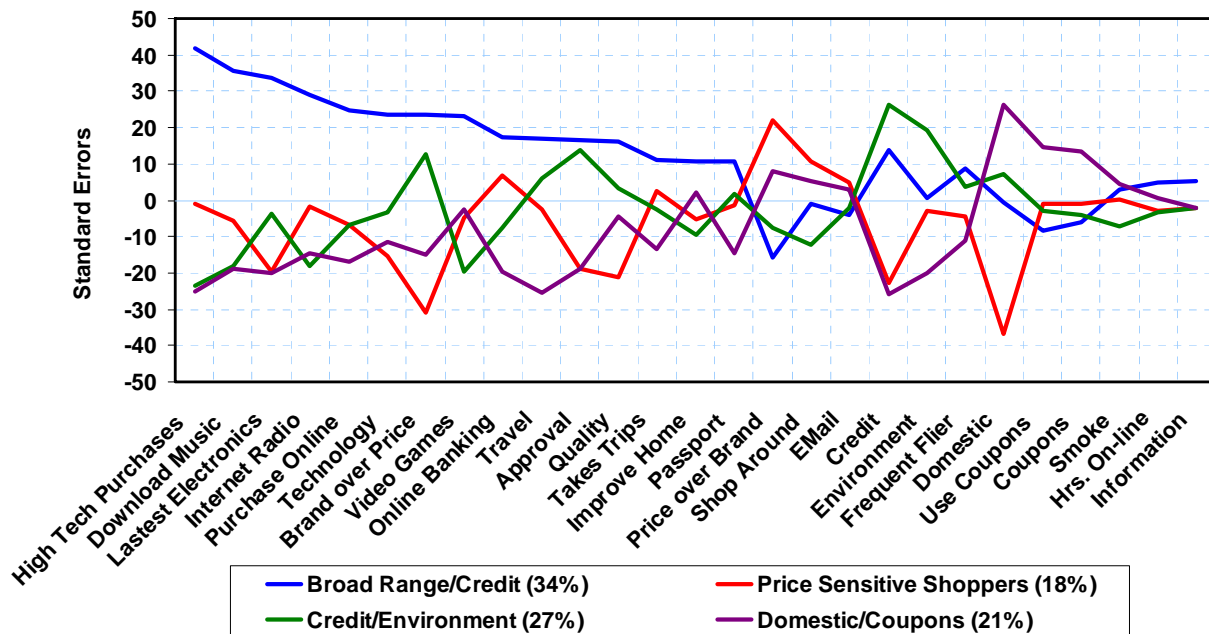


Figure 1, Standard Profiles for Buyer Behavior Segments

⁷ For the three segmentation schemes, the resulting models were totally (100%) able to reproduce the assignments.

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 segments vary widely between different countries as shown in Figure 2 below. These are likely due to cultural variations. However, we expect the distribution of these segments among panel and sources of data within regions to be far more consistent.

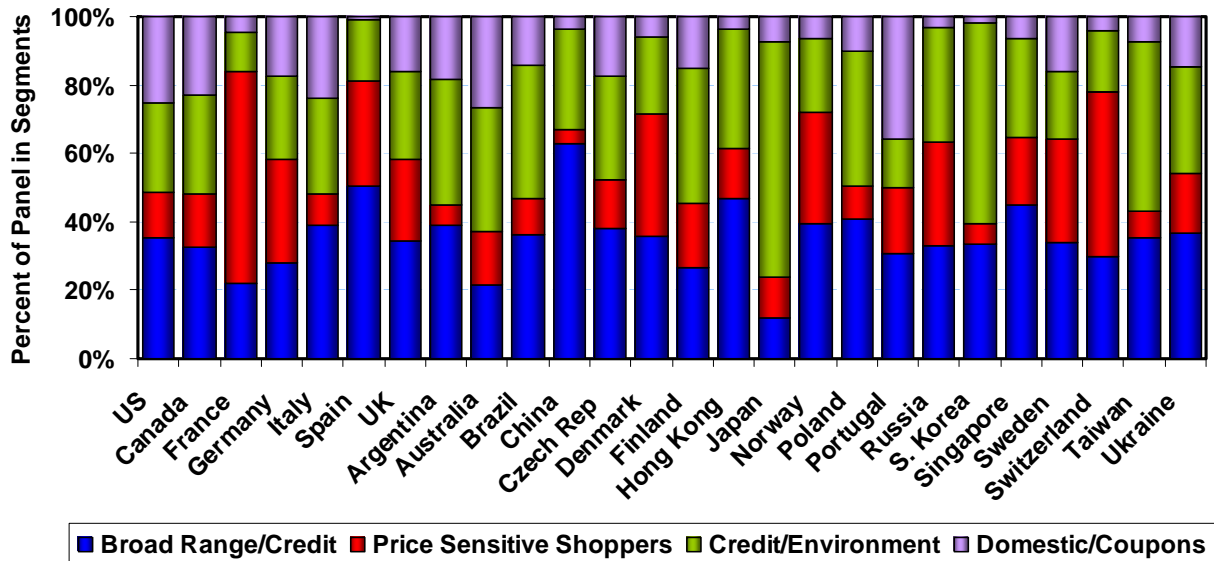


Figure 2, Country Average Buyer Behavior Segment Distribution

Figure 3 compares buyer behavior segment distributions for a typical online access panel in the United States as compared to the U.S. Grand Total. These distributions are different⁸. The largest difference is a 23% relative change within the Domestic/Coupons segment. Note that the average expected variation is 4.2%.

⁸ Based on a χ^2 test, there is less than a 49% chance that the distribution of buyer behavior segments for the panel is the same as for the U.S. Grand Total.

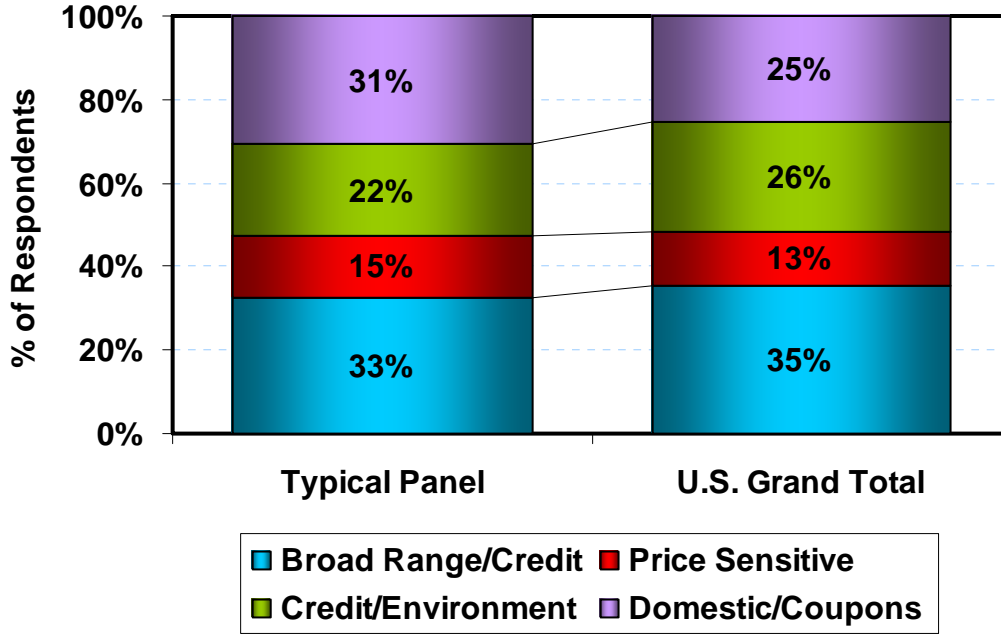


Figure 3, Comparison of a Panel to Grand Total for Buyer Behavior Segments

2.2. SOCIOGRAPHIC SEGMENTS

The sociographic segments are intended to capture the variability in behavior and attitudes regarding a broad range of life style decisions. The standardized profiles are shown below in Figure 4 and reflect the response to 31 input variables.

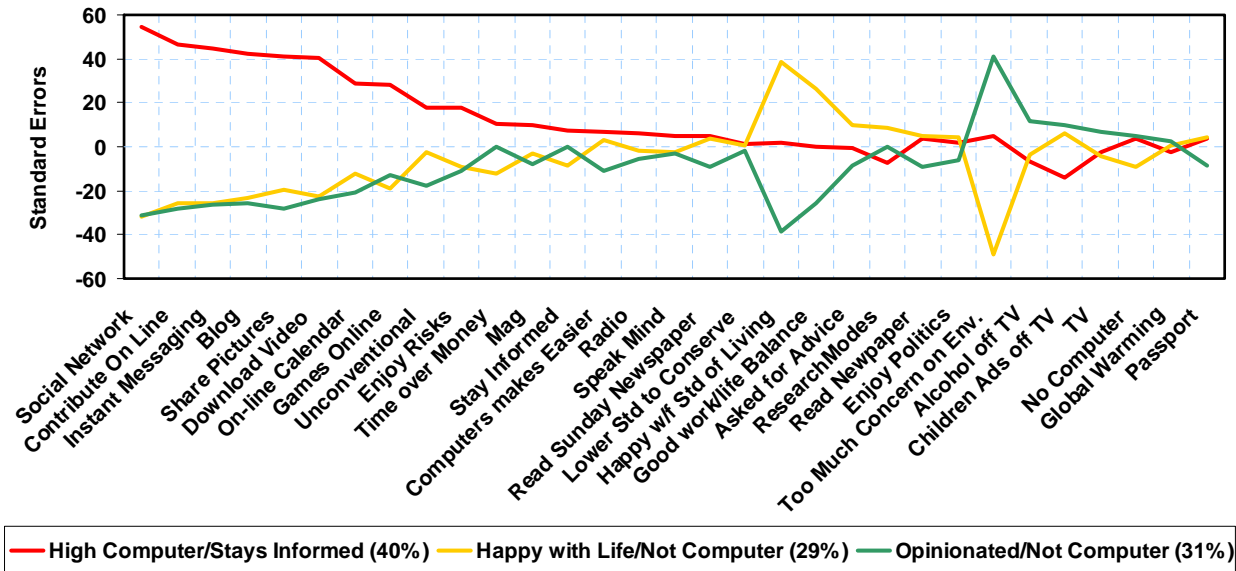


Figure 4, Standard Profiles for Sociographic Segments

As in the case of the buyer behavior segments, the titles of the segments reflect the strongest loading variables making up the segment. It is important to note that the distributions of these segments are expected to vary widely between different countries and regions as shown in Figure 5 below. However, we expect the distribution of these segments among panel and sources of data within regions to be consistent as before.

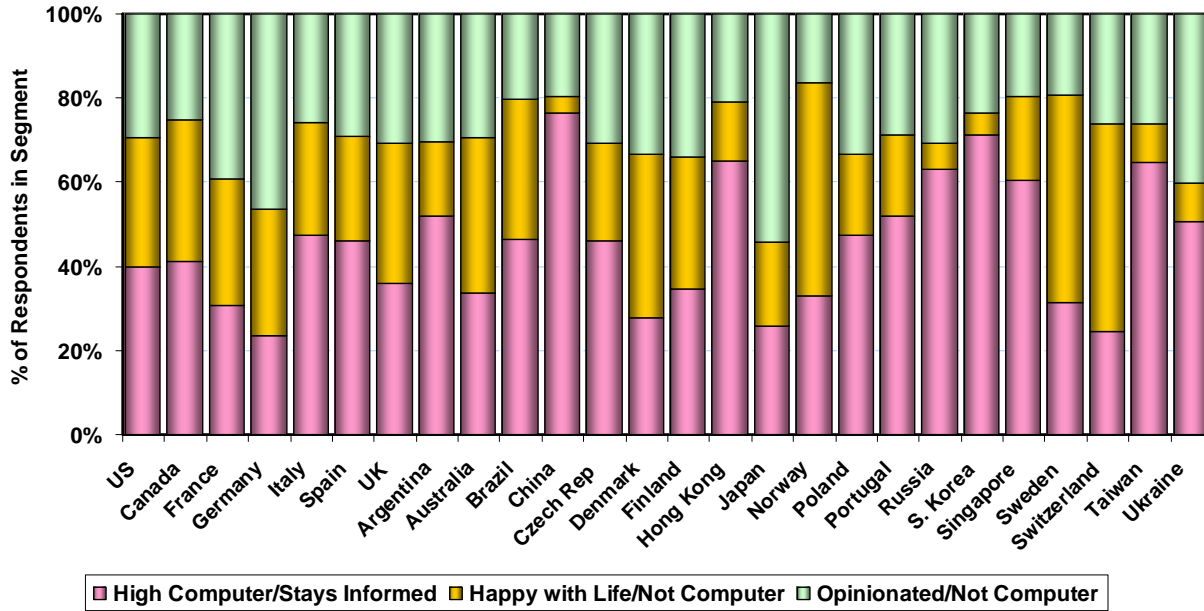


Figure 5, Country Average Sociographic Segment Distribution

Below in Figure 6 is the comparison between the sociographic segment distributions for the same typical online panel in the United States compared to the U.S. Grand Total. These distributions are also different⁹. The largest difference is 15% relative change with the Happy with Life/Not Computer segment. Note that the average expected variation is 4.7%.

⁹ Based on χ^2 test, there is less than a 53% chance that the distribution of sociographic segments for the panel is the same as for the U.S. Grand Total.

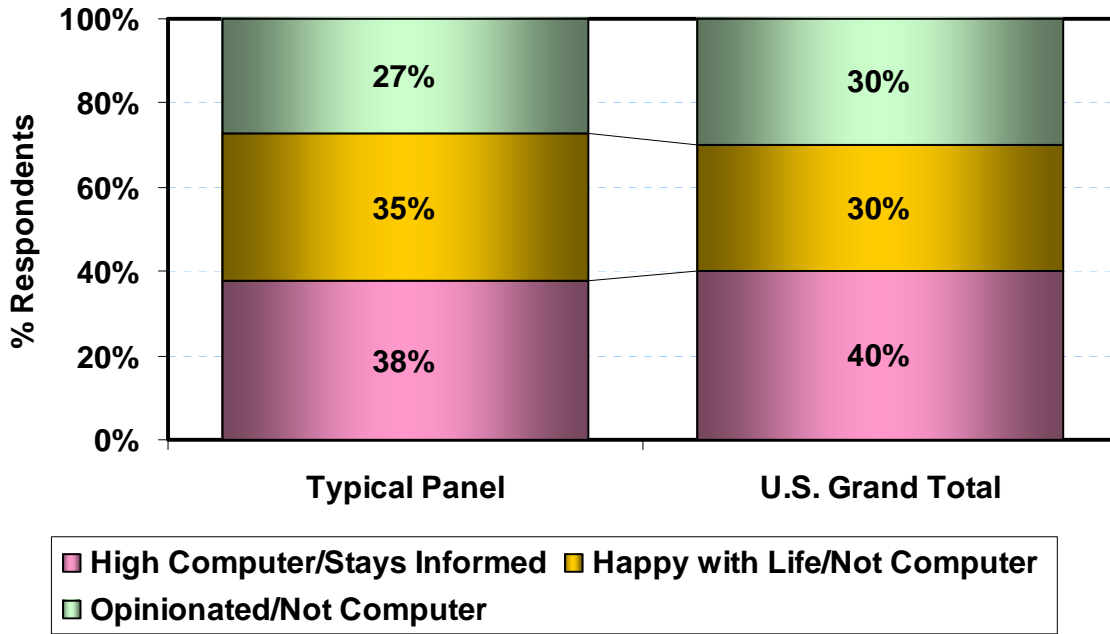


Figure 6, Comparison of a Panel to Grand Total for Sociographic Segments

2.3. MEDIA SEGMENTS

The media segments are intended to capture the variability in the use of various sources of communications and activities. The standardized profiles are shown in Figure 7, and reflect the response to 31 input variables. The variables used were combinations of those also used for the buyer behavior and sociographic segmentation but focused on media issues only. As in the case of the other segmentation schemes, the titles of the segments reflect the strongest loading variables making up the segment.

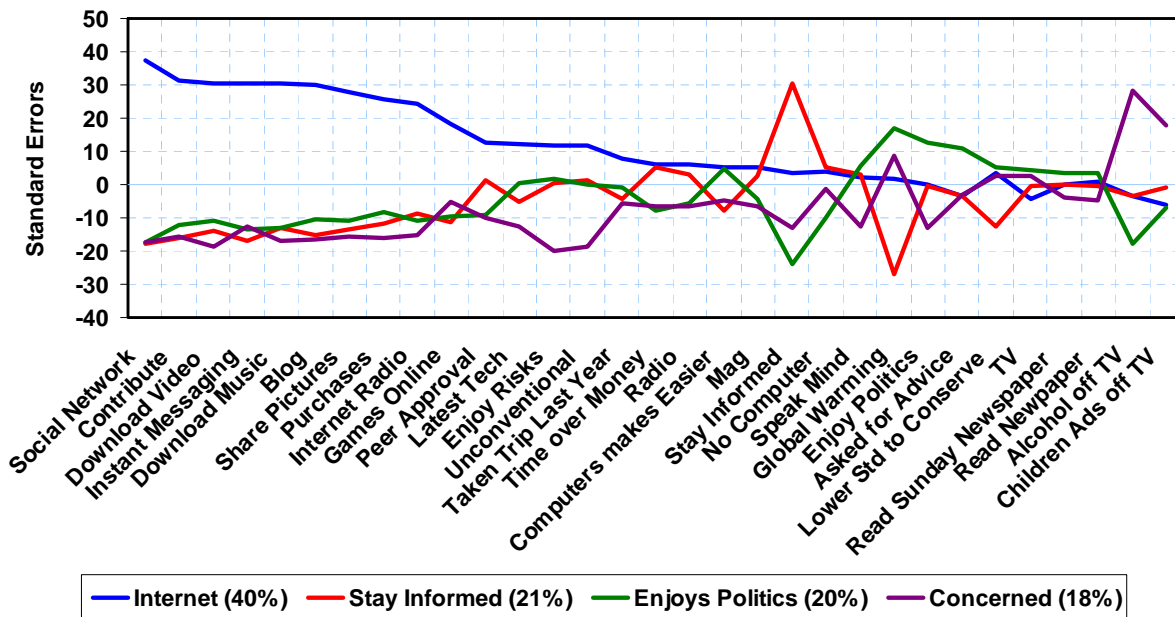


Figure 7, Standardized Profiles of Media Segments

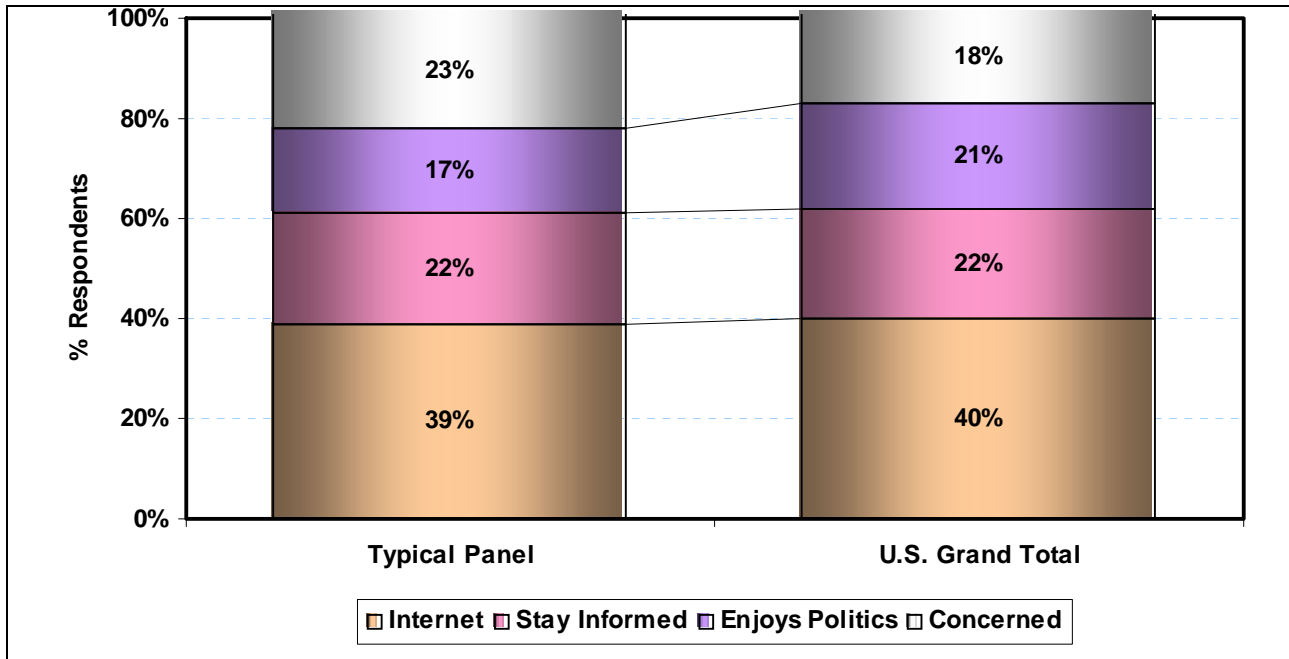


Figure 8 is the comparison between the media segment distributions for the same panel as in the above sections compared to the U.S. Grand Total. These distributions are also different from each other¹⁰. The largest difference is 20% relative change with the Enjoys Politics segment. Note that the average expected variation is 4.2%.

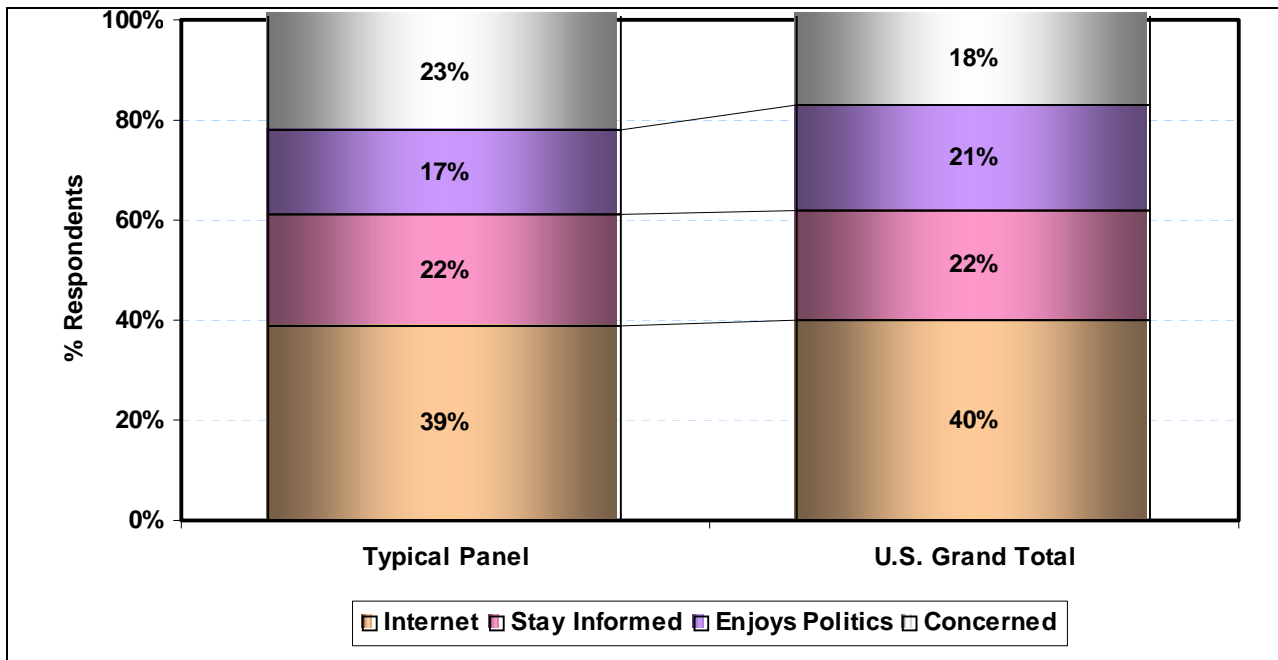


Figure 8, Comparison of a Panel to Grand Total for Media Segments

¹⁰ Based on a χ^2 test; there is less than a 65% chance that the distribution of media segments for the panel is the same as for the U.S. Grand Total.

3. OPTIMIZING OF PANEL MIXES

The causes of the disparity between the segment distributions of a panel or data source and that of the collective total for a country are probably the product of concentrations of various behavioral groups. These are in turn created by different methods of recruiting participants, the incentive process and panel management. In any event, these disparities are going to exist for any specific sample source but particularly for non-probabilistic online sources. The issue is how these disparities can be reduced from a sampling frame. Mixing of panels has been suggested as an obvious means. The result of an average would always be better than the disparity of the most deviant panel used. But there is no assurance that simple combining of panels will produce the desired result. However, having obtained metrics on the distribution of segments in the panels, optimum mixes of panels can be obtained which minimizes the disparities.

Optimization

The optimum mix is determined by varying the weight of each of the panels in a set. While any number of panels can be used for the set, usually it is desirable to keep the number small. For Buyer Behavior segment distributions there are four segments but only three are independent values since the total weights must equal one. For this case, three panel sets are used. Note, however, that the eventual optimum may have less than three if any of the weights are close to zero.

The optimum is the point where the disparity of the distribution of segments from the *Grand Mean* is at a minimum. This disparity is the objective of the optimization and is assumed to be the Root Mean Squared Distance¹¹. The size of the segments of the resulting mixed panel is the weighted average of the component panel values.

The optimum can be found graphically as shown on the following chart (Figure 9) for a particular case. The bottom-most point on the surface represents the optimum value. This point can be seen clearer on the second chart (Figure 10), which is a contour map. The red zone area represents the lowest point, the center of which would be the optimum. For simplicity, optimization can be and was done using Microsoft Solver in Excel.

¹¹ This is the square root of the sum of the squares of the difference between the panels' segment sizes minus that of the Grand Mean. This is also referred to as the Euclidean distance.

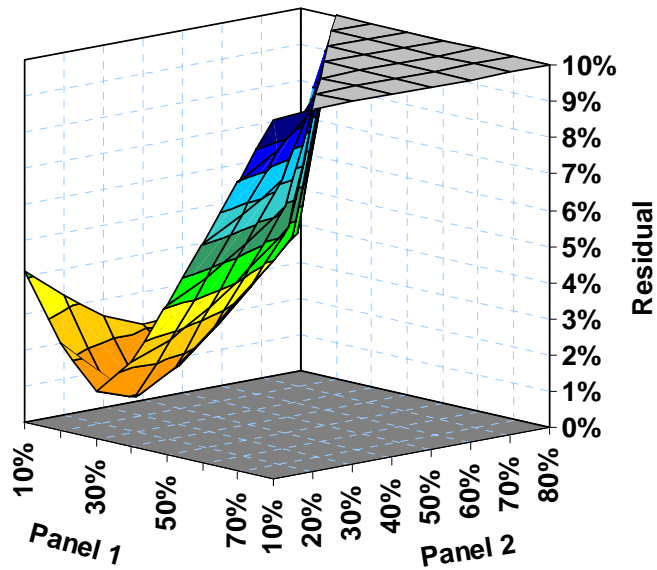


Figure 9, Seeking the Optimum, Surface Map

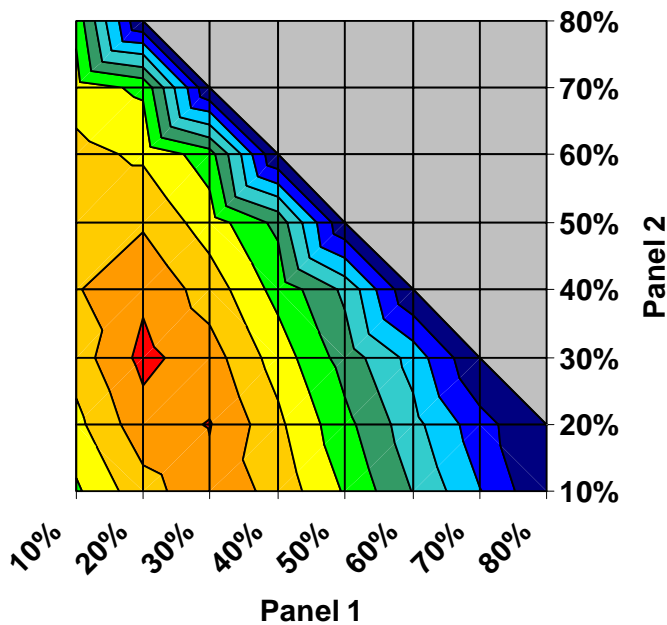


Figure 10, Seeking the Optimum, Contour Map

The Ensemble of Optima

An optimum solution was obtained from each of the sets of three panels or sources from our database of 17 panels. The Grand Mean was estimated from these data sources. This resulted in an ensemble of 680 possible optimum sets¹². The distribution of the optima is shown on Figure 11. Over 25% of these optima indicated less than a one percent deviation from the Grand Mean. However, in some cases the deviations were quite high. Not all mixes of panels can develop improved optima. In some cases two or even one panel remained in the optimum solution.

¹² This is the results of all combinations of 17 items taken 3 at a time.

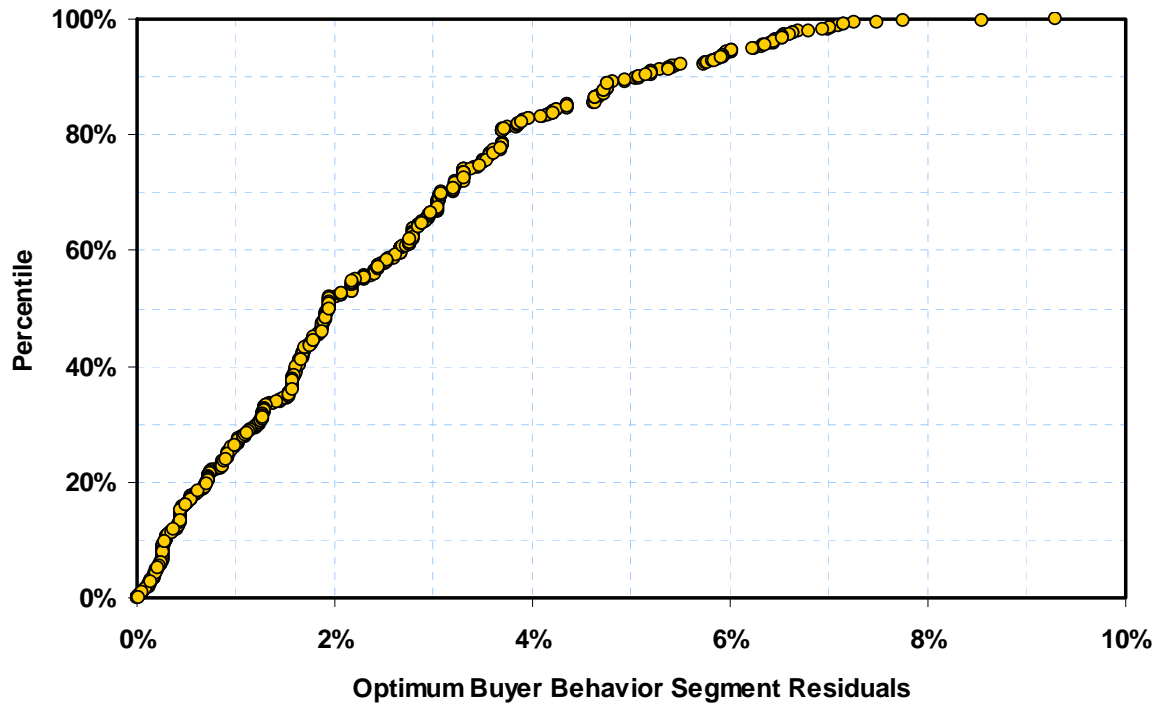


Figure 11, Distribution of the Optimum Buyer Behavior Segment Residuals

Optimum panel mixes are, therefore, not created equal. The choice of panels for an optimum mix must therefore depend on the results and other conditions. These include not only the impact on balancing one set of segments, in this case the Buyer Behavior segments, but potential on other segmentation schemes. The chart in Figure 12, shows the relationship between the residuals for the Buyer Behavior segment distributions for optimum panel sets compared to the residuals for the Sociographic segment distributions. The optimization was based on the Buyer Behavior and therefore, we would expect that the residuals will be much smaller for them than for the Sociographic segments. In general, that is true. What is important to note here is that this relationship is not monotonic, though there is a significant relationship between them. Optimizing on one set of segmentation schemes does not necessarily assure low deviations in others.

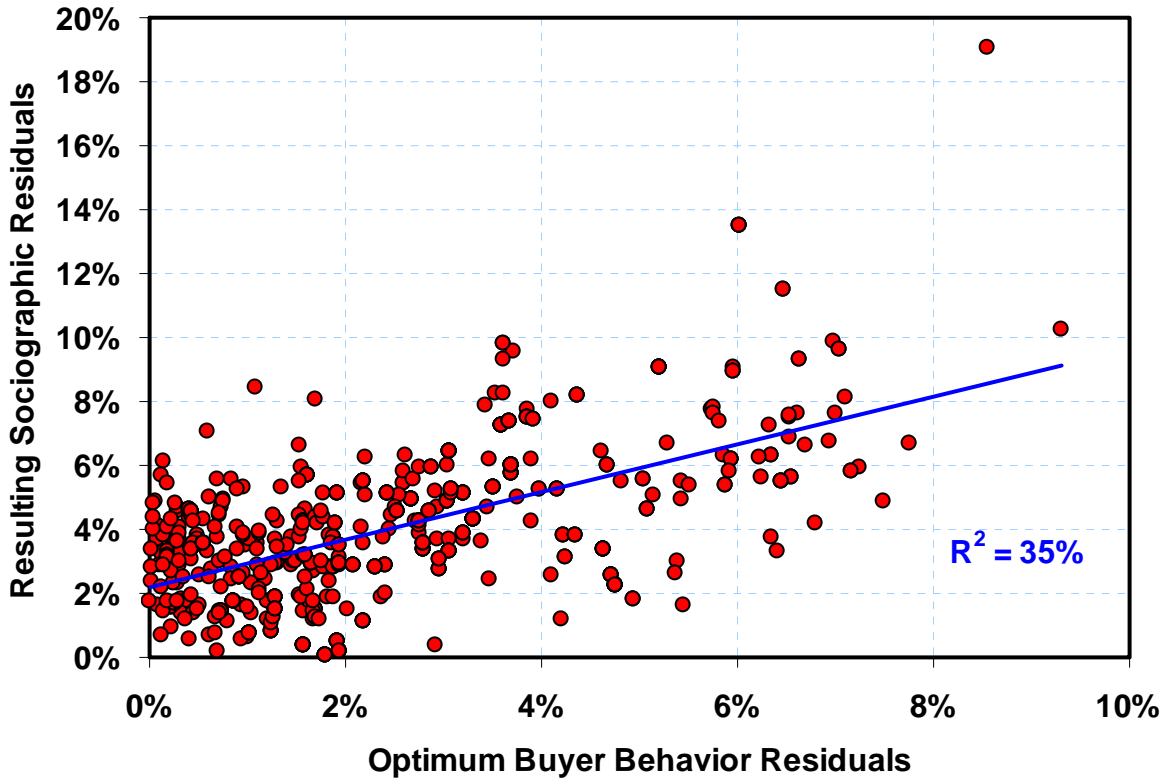


Figure 12, Residuals of the Sociographic and the Optimum Buyer Behavior Segments

4. CONCLUSIONS

Previously panel sources have been thought of as enigmatic masses of respondents that float in their own world, undefined and mysterious. This paper is an attempt to sort out mysteries of the consistency and quality of panels and to anchor the online sample universe. We do this by using combinations of quantifiable metrics, segmentation, and comparisons to the *Grand Mean* of the sample universe.

We find that we can make sensible comparisons and by virtue of structural segmentations capture the nature of the underlying sampling population. We can compare panels to one another, evaluate them through time and make statements about their consistency, predictability, and reliability.

It is the measure of sampling frames consisting of the online panels and data-sources through time that we believe is critical to establish a consistent, predictable and reliable sample frame. We call this new study, “*Consistency Analysis*” and seek to anchor online research in the virtues of this new framework. By measuring the constituents of panel samples we can better understand how they may drift through time in a world of ever changing sources. By frequent measurement and re-calibration, we believe consistency measures are possible and pragmatic. Using these tools, we must collect the data in a regimented fashion always keeping an eye on the drift from our behavioral targets.



Steven Gittelman, Ph.D.
President

Phone (631) 277-7000

Cell (631) 466-6604

Steve@MktgInc.Com

www.mktginc.com

200 Carleton Avenue, East Islip, NY 11730

About the presenter:

Steven H. Gittelman, Ph.D. has been the president of Mktg, Inc. since 1979. He received his doctorate with a specialty in optimization theory in 1976 from the University of Connecticut. He has a background in statistics and sampling theory. He is the author of many technical papers and two books.