A new *representative* standard for online research: conquering the challenge of the dirty little "r" word.

A Collaborative endeavor between Opinionology, Mktg,Inc. and Phoenix Marketing International

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

Bob Fawson, Opinionology

Introduction:

Online quantitative research accounts for more than one fifth of global MR spend (ESOMAR 2010). As a result of rising popularity, researchers sourcing respondents online are using an increasingly congested resource. Mixing sample sources has become the most common way to meet otherwise binding operational constraints. In fact, mixing may be the most common activity in online data collection: mixing sources to create a panel, mixing panels to create a sample, or mixing non-panel sources to extend reach and reduce cost. However, mixing fundamentally changes the sampling frame a study draws from; and a changing frame changes estimators. Mixing means that changing estimators don't always mean changing population parameters.

Two common solutions have been implemented in an attempt to stabilize online sampling frames: demographic quota sampling and "probabilistically" recruited panels. While these methods provide some structure to online data collection efforts, they have significant limitations. Demographics do not account for enough cross-source noise in the data to produce consistent estimators when sources are mixed, while "probabilistically" recruited panels can present cost and feasibility challenges.

We propose a third solution: *scientific source blending to a new standard*. Blending substantively differs from mixing: it addresses scientific needs in addition to operational needs. The scientific method requires that we can reproduce an experiment. While mixing is typically based on the exigencies of any given project, blending – by definition – is transparent and reproducible.

We are adrift without standards. The best blending methods emulate a standard. Here we propose a means of profiling respondents around behavioral segmentations gleaned through pre-profiling and where required, as in social networks, in real-time. At its core, it is constructed by merging the data of many modes—by application it can be employed to blend multiple modes. No such approach achieves its true value unless the resultant sample frame is representative, repeatable, and accurate.

In collaboration with Opinionology, who provided the online sample and telephone interviewing; Phoenix Marketing International, who provided the test questionnaire and environment; Mktg, Inc. applied new multi-mode models that are demographically and behaviorally grounded. This new sampling standard is now in place in the United Kingdom, United States and Canada.

Its strengths come from a common language built of the segments whose distribution has been the careful construction of heavily nested samples taken in all three countries. A combination of telephone, river, social network and panel has been used to establish the distribution of behavioral segments by demographic cell in each country.

This plan is rather robust as it allows the blending of almost all standard modes of quantitative data collection, not limited by telephone, face to face, online or mail. Further, as the modeling methods are transportable, they can become the core of sampling frames in countries with low online penetrations.

As an example, in India, where Internet penetration is some 20% and growing quickly, model based sampling standards, such as those described here, can be constructed to allow online researchers to sample in a representative and a consistent fashion.

The Challenge:

Online research is being haunted by the "r" word. Sample users want to know what the samples they employ represent. Whereas a "don't ask don't tell" state of mind has dominated the industry for the past decade the drumbeat has become loud and clear. End users want to understand if the changes that they see in their data are real or the artifacts of shifts in the underlying sample frame. In a phrase if you don't think so, then it is time to "re-think" it.

The AAPOR report on online research barely squeaked out a statement in support of the use of non-probabilistic online panels. Overwhelming concerns regarding non-response bias, the variability of the behaviors represented by the panels themselves, both within and between the panels has created a crisis of confidence. Efforts to fix the problem by treating the quality of individual respondents have not addressed the more critical issues of bias within the sample frame.

There is no safe data collection mode. Telephone samples are haunted by increasing refusal rates, wire cutters and do not call lists. Among the do not tell secrets of the industry are that most telephone studies do not address the issue of households that are cell phone only or cell phone mainly. The truth is that the spiraling costs of accessing wire cutters has fast made the sample frame weakness of online research more tolerable.

Perhaps the greatest threat to the industry is that we have a generation of practitioners who fail to understand that probabilistic sampling as represented by random digit dialed samples performed by telephone a decade or so ago were underlain by a "net" of theory that seemed to run on its own power.

Online samples are as fluid as the sourcing from which they arise. The differences between panels can be magnified by management policy, the activity of respondents, respondent tenure, incentives, just to name a few. The industry has only recently awakened to the variability of these samples and the undertow of the revelations has eroded confidence in our data.

There was a day when our samples were rooted in comparisons to the census. We could always compare the subsample we were using to assure ourselves that the sampling frame represented the census and thus the population at large. Such samples were considered to be probabilistic in nature and from there we were free to sample with the a priori belief that we were approaching a representative sample. In a sense, the census was a model that we replicated, not only by a sampling effort with a probabilistic core but often with demographic quotas that provided us with some reassurance that our efforts were on target. But what model can we fall upon today?

Online respondents behave differently within demographic groups. Panel respondents are quite different from social network respondents and the panels have been shown to differ themselves.

Thus we need new models that include both demography *and* behavior.

We need a standard that serves our purpose:

There is no perfect standard any longer, even the census. Unfortunately, for the research community behavior and demography have become uncoupled. The census provides us with a language of demography while we are in need of behavioral standards. Perhaps we should accept the census as a standard and couple it with a standard of our own design; one that works better for all of us. If, as described by the Advertising Research Foundation (Rubinson, et. al. 2009), "panels are not interchangeable," differences that exist in the behaviors that we find *within* demographic cells can be source related. This drives us to a necessary change in paradigm. We must take a lesson from our past and create a new, *workable*, set of sampling frame standards, only now because we are working without a probabilistic net we need a standard that is behavioral and demographic.

A multi-mode demographic/behavioral census:

Our industry is accustomed to categorizing behavior through structural segmentations. Thus, the combination of demographically balanced samples from different modes that provide us with a census of behavioral segments could assist us in inventing the standard that we seek.

This brings up an interesting point of concession. The perfect probability sample has become a theoretical dream. Instead, we live in a world of "fit for purpose." In essence, to move forward, we must accept a bit of dissonance in our thinking. We need to be willing to leave some things unresolved. Our standards will have tremendous pressure applied to them right from their very creation. The United States Census is not replicated throughout the world and yet we need a global standard. The end users, who purchase our research, are most often global in scope and will seek standards of global reach. There is no one-size-fits-all approach and thus we argue for latitude in our creation. In fact, we seek the right to evolve as the fast changing world continues to rotate.

We propose that respondents should be behaviorally pre-profiled and classified by a battery of segmentations that we distribute according to a demographic/behavioral standard. To this effort, we use a battery of ten segmentations in our scheme that include three general segments: buying behavior (37 variables), socio-graphics (31 variables), media (31variables), and seven market segments: automotive, appliance, consumer electronics, clothing, grocery, entertainment, insurance/banking. (When required, such as in social network and river sampling methods, a real-time abbreviated segmentation scheme can be employed. This comes at a cost; the fewer the input variables, the less stable the segmentation.)

We use a seventeen minute questionnaire designed to generate the ten segmentations geared to be sensitive measures of sample source change. The questionnaire has been translated into the languages needed to execute in 35 countries and we have obtained the cooperation of some 200 panels around the world. As an example, let's focus on survey research in the US. First we have to create a standard.

We employ telephone and online sources as our modes of choice. Online sample was further divided into three sample segments: river, social network and opt-in panel. Heavily quota nested samples of the four sources are behaviorally different (Figure 1). Telephone interviewing, completed by Mktg., Inc. and Opinionology, utilized a modified overlap design to include the appropriate proportion of cell phone only and cell phone dominant households.

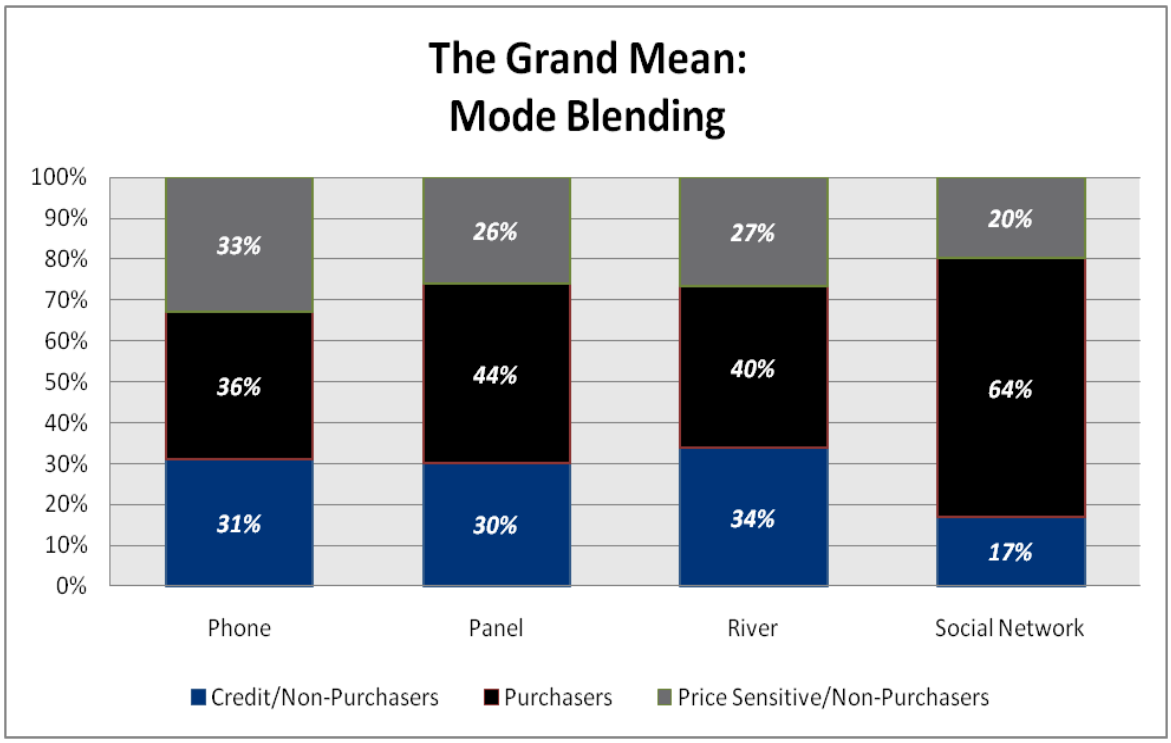

Figure 1: Grand Mean Standard™

The weights we give to samples from different modes are based upon an optimization model that seeks to minimize the difference from a battery of reference points (Figure 2). Presumably the resulting optimum combination grounds our standard to the real world. Although we carefully

collected data in the United States by phone, online panel, online river and social network, the latter two ended up being excluded in the optimal solution resulting in a 59% phone and 41% panel combination in our resultant standard. As this standard is the combination of modes, we call it the "Grand Mean Standard™" or "GMS".

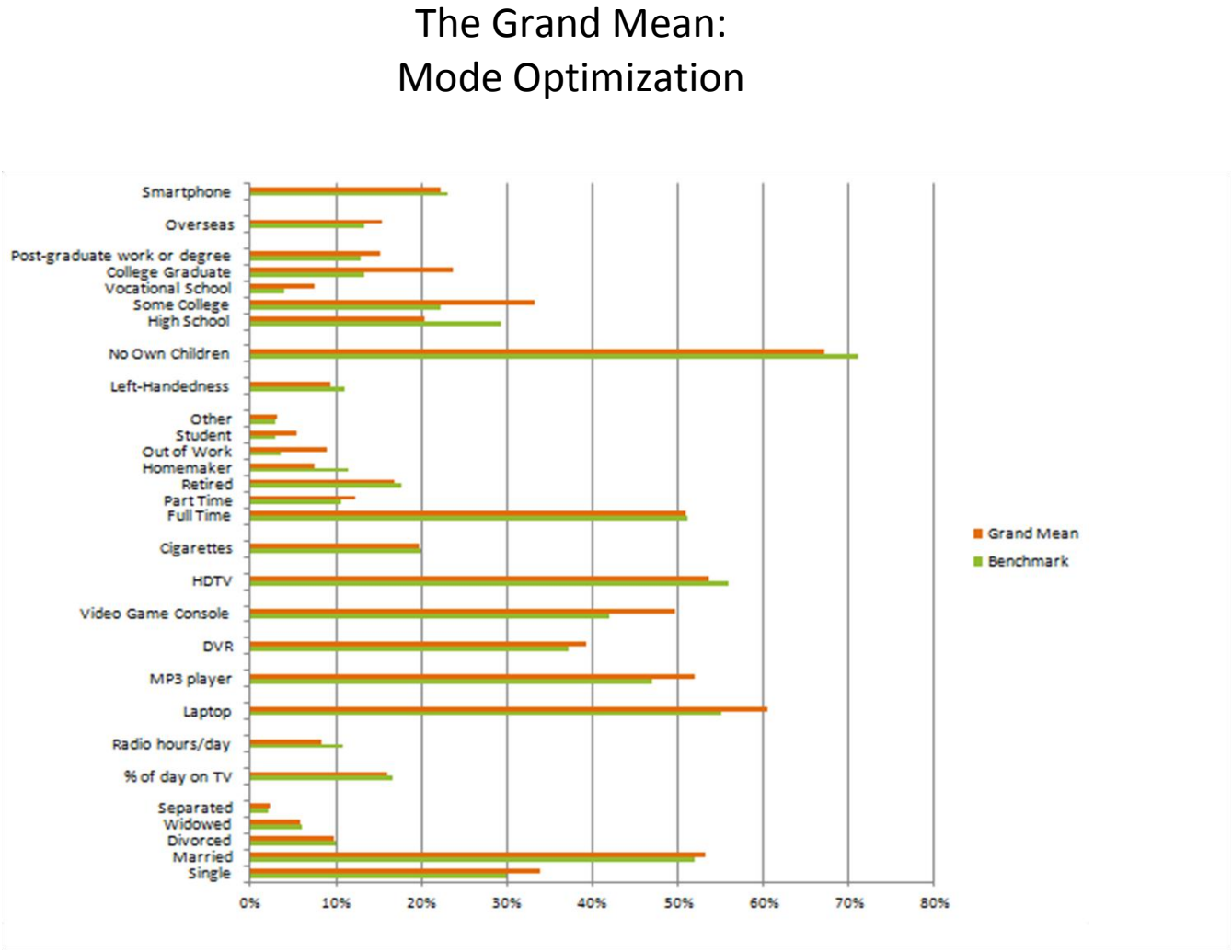

Figure 2: Fit of the Grand Mean Standard to various benchmarks after optimization.

Thus, within each demographic cell we assess the distribution of behavioral segments in our new standard. Our sampling schema is to match the distribution of behaviors in our samples to the distribution within the standard.

A Test:

In cooperation with Phoenix Marketing International (Boston) we collected data in parallel with two other companies as part of the baseline for a syndicated study.

Data was collected during January 2011, sample was provided by Opinionology (1968 completes) and ClearVoice (80 completes), for a total of 2048, to fulfill quotas. A third company, GMI completed a separate sample of 1005 completes providing a total of 5565 respondents for the study. Another company providing sample completed 2512 interviews and is known for their sophistication in sample management will remain anonymous and shall be called the Panel A.

The survey was programmed and hosted by Research Results.

Quotas were applied for income, age and gender. All companies were urged to use best practices as this was known to be the baseline of a rather public syndicated product in the financial services industry. Panel A used a platform which they promote heavily. Mktg, Inc. used a platform that is behaviorally balanced against our US standard, Real ID®. GMI employed their newly announced Pinnacle™ sampling method.

Results:

All companies adhered to quotas rigorously (figure 3). Phoenix Marketing International recommended a battery of metrics that were of proprietary concern. We compared the results gathered from the three samples on these metrics and depict them in figure 4.

PMI Payment Study: Quota Completion

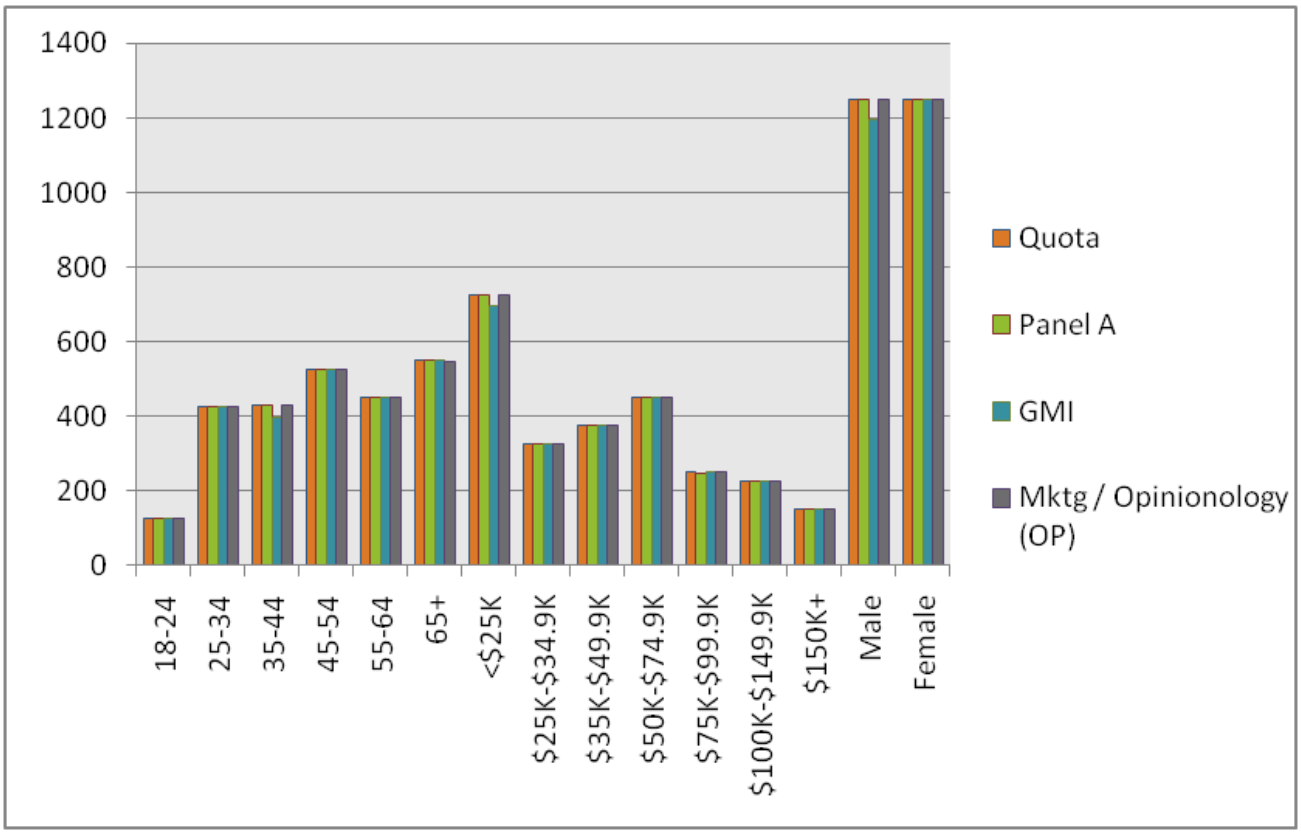

Figure 3. All samples adhered to quotas rigorously.

PMI Payment Study: Survey Metrics of Interest (Metrics provided by PMI)

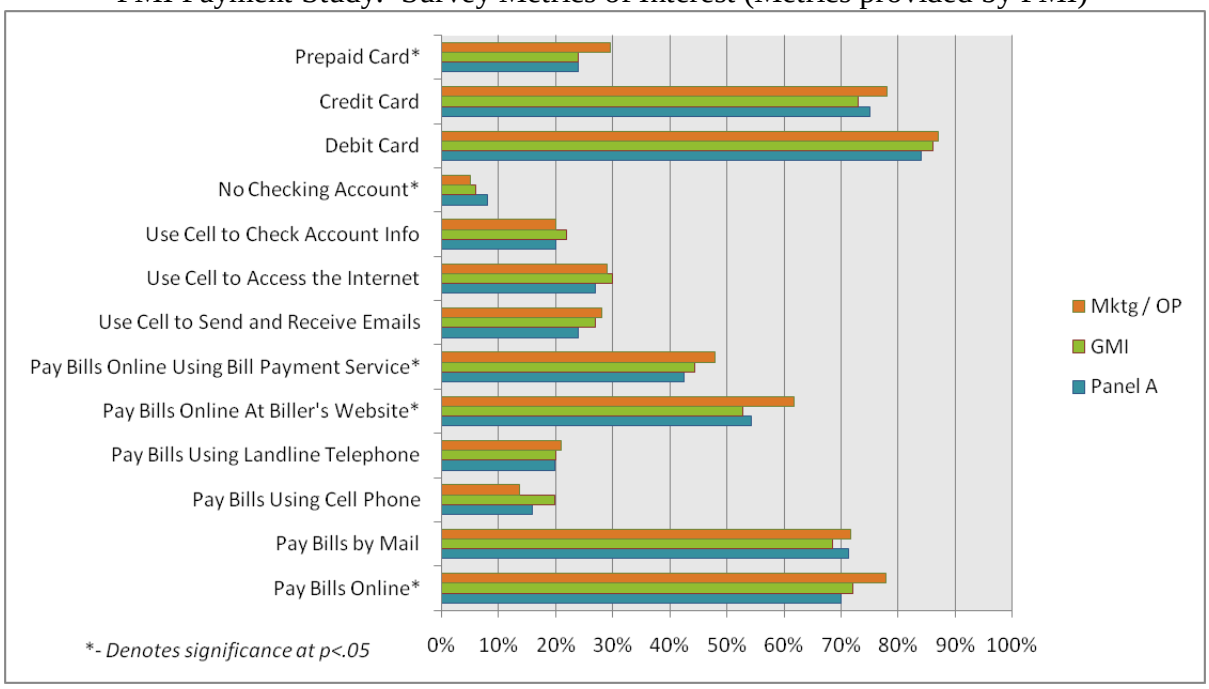

Figure 4. All samples provided a good fit to the survey metrics requested of the client.

Given the number of reference points and the differences between sample sources and methods of management the similarity of top line results was surprising.

To further examine the outcome we performed a segmentation analysis against normalized data. The results of these segmentations are depicted in figure 5 and again imply overall similarity of the data sets.

PMI Payments Study: Segmentations - Overall

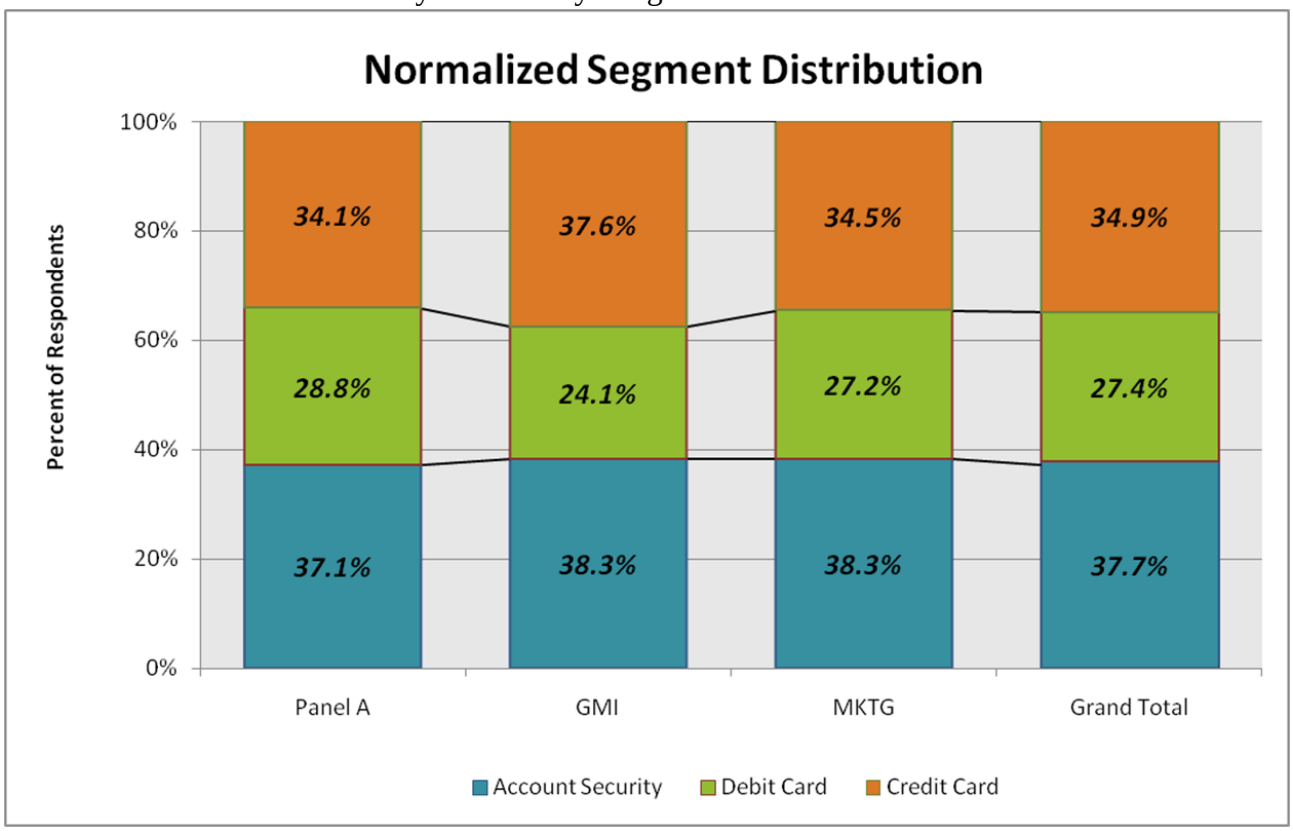

Figure 5. The results of a segmentation analysis of behavioral questions generated similar results between the three companies when looking at the overall data.

GMI employed their Pinnacle method which ties into the Current Population Survey of the University of Michigan (CPS). They use a total of sixty variables to insure that sample sources represent the results of the CPS as close as possible. They create greater precision by "trimming" in real time those variables that differ most severely from the CPS. As this is real time trimming they are restricted to a small set of variables for correction, but are able to rotate the variables being massaged sequentially so that they simulate the CPS as precisely as possible.

Real ID® is a product of Mktg, Inc. It relies on pre-profiled respondents who have been subjected to a number of quality processes the most germane here is that panelists are "behaviorally fingerprinted" through a battery of ten segmentations. Each respondent is then assigned to a segment in each of the segmentations. The distribution of segments is determined through a multi-mode study that is grounded to a battery of reference points and optimized to minimize the difference between the reference points by manipulating the percentage represented

in the final standard of each mode. In the current standard, we found that both online river and social networks had a negative effect on the fit of the modal combination against the reference points and the resultant blend is 59% telephone and 41% online panel. We call this new standard, the Grand Mean Standard™ (GMS).

As the samples gathered to create the US GMS are comprised through a highly nested quota scheme consisting of 160 cells (race x income x age x gender) the model itself is very granular. In creating a sample frame we are then able to determine the distribution of segments in each cell. In this study we used a GMS based on the buying behavior segmentation, which has a three segment solution. (Appendix show 37 questions, show segmentation) By knowing the distribution of segments within each demographic cell we are able to create a sample frame that is demographically and behaviorally distributed and in both cases represents the online and offline populations.

Because of this granularity, we balance within cell. Thus, we would expect that the results of Real ID® should not only target reference points on average but exhibit a closer fit to the distribution of those reference points then the GMI and Panel A.

Figure 6 shows the variability from reference points overall and the variability from reference points within demographic cells overall. In both cases the differences between samples sources are significant with GMI's Pinnacle providing a fit closer to Real ID® then Panel A

PMI Payments Study: Average Deviation Against Non Quota-Controlled Benchmarks

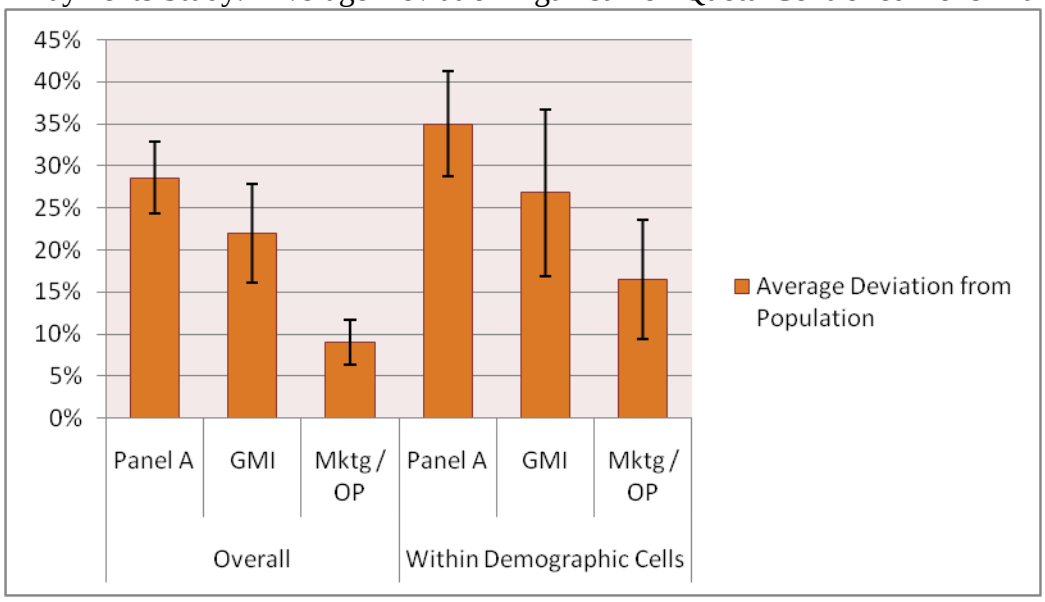

Figure 6. Deviation from population standards drawn from the Current Population Survey,was greatest in the Panel A and smallest in Real ID®. The same pattern existed when deviation from population standards was measured within demographic groups.

Conclusions:

A representative sample should not only prove accurate against reference points on average but should rise to the challenge of fitting the values of these reference points on a distributed basis. In other words being accurate on average is a massive improvement over what we might see in uncontrolled online panel samples but not actually representative of the population distribution.

However, the battery of reference points that we employed was the artifact of what was available in the client questionnaire. There is an important learning here: launching into a comparative test takes planning. We suggest that standards for such comparisons be rigorously considered before launching into the test. We need to ground our research to reference points and it appears logical to agree on a battery of such references before beginning the venture. In our case we are in search of good reference sources that are granular enough for us to make within demographic cell comparisons.

GMI did not know that we were testing Real ID® against.GMI Pinnacle They in turn were aware that they were testing against the "Panel A." "Panel A" was creating the base wave for an important syndicated study and should have been employing best practices. Nonetheless, we applaud GMI for their unfaltering willingness to have this data presented. Their transparency is refreshing. Similarly, Opinionology and Clear Voice provided the sample that we used for this Real ID®test. Their willingness to expose themselves to criticism if our methods did not work is also a refreshing breath of transparent air as well as a reminder of their commitment to online quality.

The need for testing methods is critical for building credibility and improving the concept of sample frame management. Perhaps the most salient result of this exercise is our belief that sampling standards should be elevated. One way of accomplishing this task is for blind comparative tests to be sponsored by sample users. It is our hope that full service companies will adopt this practice and that end-users will appreciate the effort and expense that goes into achieving the best sample frame that we can create.

References:

Baker, et.al., March 2010. AAPOR Report on Online Panels**.** Prepared for the AAPOR Executive Council by a Task Force operating under the auspices of the AAPOR Standards Committee.

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.

ESOMAR, 2010. Global Market Research 2010: ESOMAR industry report.

Biographies:

Steven H. Gittelman, Ph.D Mktg., Inc. - President

Steve Gittelman and partner Elaine Trimarchi have championed the need for standardized metrics for online research. Their company Mktg, Inc. has spawned a new division, Sample Source Auditors™, now known best for the continuous tracking study, The Grand Mean Project® and Consistent Track®. Both have thirty years of data collection experience and have made the pursuit of online quality their mission. They have rigorously analyzed over 200 global sample sources in 35 countries. In their view, we should be able to identify real shifts in data from variation in the sample frame.

Bob Fawson, Opinionology - Vice President, Online Services

Bob is Vice President of Online Services at Opinionology. Bob has leveraged his education, industry experience, and research background to grow Opinionology's portfolio of online sampling methods and improve methods to intelligently manage online panel inventories. Bob continues to explore the nuance of online research methods and panel management techniques. His efforts have improved industry understanding of many diverse topics from panel recruiting and sample routing to address based sampling. He is often asked to present his findings in market research forums. Previously, Bob managed Opinionology's online client service and project management team. Bob holds an MS in Political Science and an MS in Applied Economics from Utah State University. He joined Opinionology in 2007.