On Probability Sampling, Babies and Bathwater

(A Media Rating Council Staff Point of View)

Discussions of samples used for Internet panel (user-centric) research invariably turn to the viability of probability sampling. Because "the Internet is different," it is often stated that traditional media research principles of pure probability sampling must be set-aside to make large Internet-based panels, driven by the need for visibility into large numbers of measured web-properties, practicable.

That, of course, poses potential problems for the Media Rating Council (MRC) where probability sampling is implicit in some of our published *Minimum Standards*. The MRC now has to ask itself whether alternatives to probability sampling *could* be acceptable under certain circumstances, and if so, what those circumstances might be. In considering those questions, the MRC Staff believes that Internet-panel issues go well beyond the textbook challenges of probability sampling, and that the challenges must be considered more discretely.

The ultimate quality of a sample depends upon many factors:[1]

- Knowledge of the universe
- Coverage of that universe (e.g., minimizing totally excluded populations)
- Efforts to minimize nonresponse bias in general (e.g., optimizing cooperation)
- Understanding and minimizing differential cooperation (to achieve proportional samples)
- Consistency of procedures over time (so that results are replicable)
- And sample sizes, of course (for point-in-time precision)

While we believe it is fair to take a fresh look at issues of probability sampling, it is also essential that other sampling "best practices" not get lost in the process. In short, we need to ensure that certain critical quality principles (the "babies") are maintained despite conclusions we may reach on probability sampling (the potential "bathwater")…and we need to have a logical process for asking the *right* questions about probability sampling.

Having defined the larger scope of the issues, we will analyze them one at a time in this paper. Thereafter we provide our thinking on how to determine the necessity of the MRC's probability sampling principles in the Internet measurement environment.

[1] In this write-up, we're only discussing sampling issues. Of course, there are many other operational and disclosure-related Standards issues which receive extensive consideration at the MRC for all forms of research, including Internet-related.

Universe Knowledge

We suspect that Internet measurement practitioners would agree that media currency measurement needs a high-quality assessment of the broad parameters of the universe. Whatever usage is ultimately measured, and however it is defined, the Industry needs a reasonably bullet-proof assessment of key measures of the Internet *population*.

Reasonable people will disagree on the frequency of benchmark measures. Similarly, there's *some* room for disagreement on just how high-quality that benchmark has to be. But we believe there is general Industry agreement that benchmark data from a sample that is fully defensible is a necessity. Today, that argues for probability sampling and reasonable response rates *for a periodic universe (benchmark) survey*. We assume there is agreement on that principle, so our focus will be on the details of that survey's execution and on its frequency; and conventional MRC Standards directly apply.

We will come back to this point, but in essence: If large, less-scientifically selected samples are to be adjusted to something, then that "something" must at least be of reasonably high quality. The MRC *Minimum Standards* are up to the task of facilitating assessment of this procedure.

The Sample's Coverage of the Universe

Internet panels have been known to exclude college students living away from home, businesses of certain characteristics, and even Internet users beyond those using certain sites (those used for recruiting) as examples. The hope, of course, is that the users that *are* selected can become reasonable proxies for those that never had a chance to participate.

This issue goes well beyond probabilities of selection, but it is also not unique to Internet measurement. It is common to exclude nontelephone households from telephone-based surveys, certain dangerous neighborhoods from area-probability based selections, and now many surveys wrestle with the issue of cell-phone-only households. And of course, there are occasional "technical" issues or limitations in meter-based measurement environments to varying degrees. In short, we have learned to live with certain population exclusions in media research.

We do so, however, on the basis of two consistent metrics—materiality and dissimilarity. Are the excluded populations material in size, and if so, are their survey-relevant behaviors believed to be substantially different? If the answers are yes, the accuracy and MRC Standards-compliance of a currency service has to be challenged.

We owe Internet measurement the same scrutiny. Excluded populations have to be sized within reason, and reasonable assessments have to be made about their behavioral differences. Absent that knowledge, we lack a sufficient understanding of the potential for bias. Only when this issue is understood, can we consider whether the resulting potential biases can be mitigated through subsequent weighting of the *included* samples. Understanding the potential bias from excluded samples is difficult for the very reason that they go unmeasured in the first place; therefore, it's difficult to prove that bias issues related to their absence can be overcome.

The challenge before a currency-research provider with significant amounts of excluded population is to quantify and understand those populations, and to demonstrate empirically that such biases (if potentially material) can be reasonably corrected. This is not a probability sampling issue; however it is an issue critical to overall research quality of that provider.

Optimizing General Cooperation

Despair over low survey cooperation is hardly unique to Internet measurement. In fact, cooperation difficulty is one reason why practitioners came to question the value of rigorous response-rate linked sampling techniques (such as probability based sampling) in the first place; if the best most companies can do is response rates that are well below 50%, why bother trying?

A fair question…Do we really know that a 30% response rate yields a more accurate result than, say, a 5-10% response rate?

Well, no, not with *certainty*. But as Bertrand Russell said, "When one admits that nothing is certain one must, I think, also add that some things are more nearly certain than others." Assuming that higher response is better seems like the best starting assumption. One can argue, as some modern academic researchers now do, that nonresponse bias can vary from measure to measure, and survey to survey, without abandoning the principle that "good survey practices [still] dictate striving for a high response rate as an indicator of the quality of all survey estimates."[2]

To deviate from that assumption, we believe the burden of proof has to be with the research supplier. We believe that it is reasonable for a user to assume that a higher-response sample is different (i.e., better) unless a provider can prove differently. Therefore, it is reasonable for the users of media currency to expect a supplier to always strive for higher cooperation unless it can be proved convincingly that such efforts are unimportant for a particular survey with a particular set of measures.

Furthermore, it seems to us that quantifying the differences between high-cooperation and low-cooperation samples would be a prerequisite for claiming that weighting could compensate for any resulting differences. Since eliminating bias is almost always preferable to weighting compensation anyway, shouldn't the industry always expect a currency provider to measure cooperation and to treat cooperation optimization as a high priority?

But we recognize this is a significant issue with some of today's Internet samples: There is no history available for sample acquisition, and therefore, no ability to compute even a simple cooperation rate.

We believe this is one of the most important and fundamental problems facing today's Internet samples. Since we believe that optimizing cooperation is an important component of quality, we also believe that suppliers need sample recordkeeping and history to assess and act on that cooperation. Today, it appears that such recordkeeping is frequently nonexistent.

---

[2] Groves, Robert M. 2006. Nonresponse Rates and Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70:5—670.

Again, this is not a theoretical probability of sampling issue…it is a research quality issue.

Achieving Proportional Samples

Some population groups use media differently. Some of these groups are also unusually hard to reach, hard to recruit, and/or hard to maintain in a panel.

All of these factors conspire to mean that it is not enough for a provider to merely set and meet reasonable targets for overall cooperation. A currency research service must also address proper representation *and cooperation* of groups with varying media habits.

This is an area where the Internet can legitimately lay claim to important differences. The types of population groups that need proper representation for Internet research are likely quite different (and different in priority) than those receiving attention from traditional media. The specific variables needing attention (for example, form and speed of Internet access) are likely to differ from traditional media.

Assuming that the "proper variables" are known, some would argue that sample weighting can address the issue of sample representation adequately, and up to a point, it can. But we need to point out some major caveats about weighting.

Here's what Robert Groves said about weighting and nonprobability sampling in a recent edition of *Public Opinion Quarterly*:

> It is noncontroversial to note… that by departing from randomized selection, those designs burden the analyst with adjusting respondent estimates *both* for nonrandomized selection procedures *and* for nonresponse. … Whether the nonprobability sample survey can fulfill the heavier adjustment burdens is a function of what auxiliary variables are available. Nonprobability samples with explicit frames (e.g., address samples, RDD samples using quota schemes to select persons) generally have more auxiliary variables for adjustment than nonprobability designs that merely use volunteer samples.[3] [Italics added]

In short, the less the provider knows about the origins of the sample, the harder it is to make weighting arguments, or to prove that weighting is effective.

First, the supplier needs to *know what matters* for weighting. Age and gender may be necessary, but they are likely to be insufficient stand-alone. The variables appropriate for Internet weighting require careful research by an Internet currency provider.

Second, weighting does not necessarily compensate for significant *differences in cooperation*. If an important population group has half the cooperation rate of the general Internet population, that fact drives a significant increase in the potential for behavioral biases that potentially may not be corrected through weighting. The supplier needs to know the source of under-representation and address cooperation issues through differential survey procedures to reduce that potential for extra bias.

---

[3] *Ibid*., 667.

Third, there *is* such a thing as *too much* weighting. A weighted sample loses stability, and a thoughtful, empirical process is required to assess those tradeoffs.

None of the above represents probability of *selection* issues.

Consistency of Procedures and Reliability

For a variety of legitimate reasons, large sample sizes are necessary in Internet measurement.

But some of the apparent reliability gain from large samples may be a chimera if the sampling procedures are inconsistent. If the sample-source for the current survey differs significantly from the sample source in a subsequent survey period, or if a panel is refreshed from wildly varying parent lists, stability becomes unlikely, even with large numbers of respondents.

The reliability of panels and surveys is more than just whether traditional standard errors can even be computed—a technical issue raised by nonprobability sampling. The problem is that this year's sample may not be at all predictive of next year's sample if the procedures and sources vary in uncontrolled ways.

We believe sampling procedures have to be reasonably consistent and replicable for "reliability" to have *any* meaning. If a supplier can demonstrate *consistency of procedures*, the Industry might benefit from disclosure of a standard deviation computed from that sample—but not before.

Procedural consistency is not a probability issue per se…it's just a matter of quality.

Probability Sampling

Now let's assume that a currency provider has addressed all those other issues. Let's suppose:

- They create a high quality source of universe estimates, and
- They provide reasonable coverage of that universe with their sampling (and/or can make reasonable disclosures of the materiality of exclusion biases), and
- They make continuous, good-faith efforts to measure and optimize overall cooperation from those samples, and
- They understand and treat the population groups that warrant differential treatments, and then provide *appropriate* weighting for lingering imbalances, and
- They can demonstrate robust, replicable, consistent procedures over time.

BUT: The currency provider does not design and execute a probability sample in all parts of the service. What now?

If those other issues have been addressed to a reasonable degree, *then* the door is open to considering nonprobability sampling for portions of a service, we believe. For example, in principle we could live with some population members having zero chance of selection by the provider. To a degree, we already have that problem today, and not just with uncovered populations in the frame.

Let us simplify the world, and assume that one-third of the general population will never, ever participate with any non-Census survey. That's roughly a safe assumption, which means that about a third of the population already has a zero chance of being in a survey, even if they're "selected." Furthermore, we know next to nothing about them except perhaps their phone numbers or addresses.

So we already live with the fact that some people will never be in a survey. We now live without knowing much about them, though we do so today because we have no choice.

In theory, then, that makes today's standard error calculations mean less than they once did. Those error terms really only reflect the variability of surveys conducted consistently among the cooperating population. Therefore, survey or panel which met the characteristics described above could presumably have such a statistic created.

Calculating a true response rate is a little tougher. A true response rate requires knowing the denominator of "predesignated sample." Since much Internet sampling is not really based on an initial selection of persons or households, we do not have a strict number of unduplicated "persons attempted." But if we at least had records of attempted recruitments (see earlier discussion), we might be prepared to consider alternative metrics.

But again, that requires solving what we now believe is one of the biggest challenges facing most current Internet sampling—tracking the details of the recruitment process in a way that allows complete cooperation rate calculation, regardless of source. That will be a significant goal of the MRC dialogue with Internet panel measurement suppliers.

The Bottom Line

Overall, we're ready to consider some compromises in strict probability sampling in the name of larger Internet samples for currency measurement.

But we will not let certain narrow technical issues allow us to throw out the quality "baby" with the probability "bathwater." Research providers should understand that MRC will be diligent in ensuring that cooperation proportionality and consistency processes are not part of the compromises being studied related to probability sampling.

Specific Implications

To help guide Internet research services (and ourselves), the MRC Staff prepared the Analytical Guidelines provided in Appendix A.

These are the questions and analyses that we believe are:

1) Consistent with the philosophy expressed above
2) Necessary for MRC members to assess how Internet research panels might comply with MRC *Minimum Standards*
3) Appropriate for Internet research providers to disclose to knowledgeable users in conveying the true quality of their currency-level products

We recognize, of course, that Internet measurement of all types will be a dynamic and evolving area, and that we ourselves will learn along the way. We expect feedback, and we intend to be accepting of constructive suggestions.

In the meantime, however, we hope that this paper clarifies how the MRC Staff is approaching the challenging area of assessing Internet user-centric measurement panels.

Appendix A

Analytical Framework for Assessing Internet Measurement Panels

Nonprobability Panel Projection Validity:

- Jackknife replication of nonprobability panel to quantify standard deviations of frequently used measures, including estimated errors of the difference (i.e., predicted deviation of the difference between estimates from two such samples)
    - Then contrast those single-point-in-time estimates of standard deviation with standard deviations computed from the panel at two points in time, preferably at least one year apart
    - If deviations over time are significantly larger than deviations estimated from the initial replication, this could imply material inconsistency in procedures (deviation beyond that associated with sampling)
- Comparison of predicted deviations for common estimates in this panel with standard errors for common estimates for other media (e.g., television)
- Estimate the impact on reliability from conforming to Calibration Probability-Based Panel (if used), including potential impacts from alternative sample sizes

Other Corroboration – Use validated tagging techniques across selected (MRC Accredited) web properties to determine census estimates and compare these estimates, over time, to projected mega-panel estimates. Attempt to tie tagging results, where possible, to research provider panelists to assist in corroboration process. Reconcile major sources of difference, where possible.

Respondent Cooperation:

<u>Sample Frame Sources – Enumeration Process</u>

- Provide details on frame definition, including estimates of populations covered/uncovered
- Justification of Frequency of Enumeration
- Provide detailed response rate calculations for most recent year, broken down by all material types of non-contact and non-cooperation (E&Y can provide guidance on typical response rate disclosure categories).
- Based in part on the analysis above, determine operational opportunities for improvement:
    - For non-contacts, consider sufficiency of recruitment attempts, contact scheduling, in-language recruitment and interviewer management
    - For refusals, consider testing or using pre-contact mailed warm-ups and premiums, refusal conversions, incentives for completion, using refusal specialists, and/or enhanced interviewer training
- Analyze response rates by types of geography to determine opportunities for focused improvement (e.g., response rates by Metro/non-Metro, ethnic density, Census income categories, etc.).

Probability Calibration Panel (assuming recent recruitments have occurred)

- Provide details on frame definition, including estimates of populations covered/uncovered
- Provide detailed response rate calculations (through initial agreement) for most recent year, broken down by all material types of non-contact and non-cooperation (E&Y can provide guidance on typical response rate disclosure categories).
- Based in part on the analysis above, determine operational opportunities for improvement:
    - For non-contacts, consider sufficiency of dialing attempts, call scheduling, and interviewer management
    - For refusals, consider testing or using pre-contact mailed warm-ups and premiums, refusal conversions, incentives for completion, in-language recruitment, using refusal specialists, and/or enhanced interviewer training
- Analyze response rate by types of geography to determine opportunities for focused improvement (e.g., response rates by Metro/non-Metro, ethnic density, Census income categories, etc.).
- Provide details of panelist installation success rates from those agreeing at recruitment, by categories of sample (using classification questions for recruitment).
- Based in part on the analysis above, determine operational opportunities for improvement
    - Identify distinct reasons for failure to install, and propose possible solutions for each
    - Determine if certain types of sample are more likely to fail to install, and identify potential enhanced procedures for those sample types
- Turnover analysis (first steps):
    - Quantify the incidence and frequency of voluntary back-outs (to the extent possible), and evaluate how back-out incidence may vary by sample type
    - Propose possible solutions for at least the worst-case back-out categories
    - Analyze and contrast the sample characteristics of panelists installed for various tenure durations, including behavior-based variables not used for weighting

Nonprobability Panel

- Determine whether any portion of the sample used for Nonprobability Panel recruitment can have its initial cooperation estimated:
    - Can at least *some* sources of sample used for Nonprobability Panel recruitment provide data about the number of initially-attempted sample relative to the number that agreed to cooperate with the original source provider? If so, provide that data.
    - Whatever the current state of affairs, determine if research provider has future options for Nonprobability Panel sample selection that would at least allow tracking of initial cooperation (i.e. the rate of agreement to participate with the third party that provides the sample).
- Estimate the cooperation rate for initial Nonprobability Panel recruitment—what percent of initial Nonprobability Panel recruitment attempts result in agreement?
- To the extent possible, determine whether some initial-recruitment sources have lower cooperation than others, and quantify the differences
    - If some recruitment sample sources provide demographic data or other sample characteristics, estimate the Nonprobability Panel recruitment success rates by type of sample for those sources
- Identify opportunities for enhanced cooperation with recruitment overall, and for known worst case sample types

- Test or implement warm-ups, enhanced incentives, additional attempts, refusal conversions, multi-mode recruitment

Work-Panel (if separately maintained)

- Review Response Rate Calculation
- What are major drivers of response rate issues, for example:
  - IT policies which may prohibit installation of meters
  - Privacy concerns
  - Software compatibility issues?
- Exploration of alternate recruitment mechanisms
- Review Completeness of Work Panel Frame

Empirically-Based Weighting Methods:

- Develop and executive multivariate analyses of which panelist characteristics correlate most strongly with the most commonly reported measures of Internet behavior
  - Conduct a series of step-wise regressions on various Internet measures to develop the most efficient combination of variables for weighting
  - Determine whether work and home measures of Internet usage warrant separate weighting models when measured with separate panels
  - Usage of Behavioral Variables – for example, subscribers to AOL
- Determine which combination of potential weighting variables has the optimal net effect on bias reduction while minimizing increases in variance caused by weighting