White Paper

Sample Blending: 1+1 > 2

By Jackie Lorch, Vice President, Global Knowledge Management, SSI
Kristin Cavallaro, Knowledge and Data Analysis Specialist, SSI
Robert van Ossenbruggen, Consultant, SSI

Science. People. Innovention

Multi-source samples have largely been considered a "necessary evil"—a requirement when no single source can handle a large or low incidence sampling need. Historically, researchers have shied away from using multiple sources—perhaps because sampling theory dictates the use of a defined, single population. However, in a changing world where online communication is moving away from e-mail and where social media is playing an increasingly important role in opinion sharing, the traditional e-mail-driven access panel paradigm is no longer sufficient.

In fact, the single most important characteristic for an unbiased sample is that it reflect the heterogeneity of the target population. Thus, increasing heterogeneity by blending samples from many different sources, each with their own unique bias, should result in improving sample quality. But only if we are able to identify this bias in order to control for it. If bias is correctly controlled for, the use of sample from multiple sources will improve the heterogeneity of the sampling pool and thus its representativeness.

The only way to balance a multi-source sample is to get a real understanding of the source of its unique bias. One of the main problems with sampling from panels alone is that some people will never be in a panel. Just as a manufacturer of golfing equipment who conducts research solely among members of country clubs would be missing large segments of the target audiences, panels may be missing a distinct type of opinion-giver. While many people want to give their opinions, some shy away from the commitment of joining a panel. A blended approach incorporates panels, communities and groups with aligned interests. The whole Internet becomes the panel.

As well as being diverse, sample must be consistent and predictable. We must find a way to both increase diversity and deliver that diverse sample time after time. Recent research on blending has included work done by Michael Fallig and Derek Allen of GfK and the ARF's Foundations of Quality study, both in 2009, as well as others. Studies in this area have found that different sample sources have distinct characteristics, that there are many factors influencing these differences, that it is difficult to identify these factors but that controlling for demographics and other traditional balancing factors does not smooth out all of the variation. The ARF study identified recruitment techniques, (not just sources, but the recruitment approach), as well as the composition of the panel by longevity, reward methods, and other factors which haven't yet been fully researched and understood. In this paper we try to reveal the distinct characteristics of different sample sources and a way to smooth out between-source variation.

How Can a Sample Buyer Ensure a Consistent, Reliable Blend?

To ensure a consistent blend of sample, we can look at sources and see how they perform against external benchmarks, when available, but most research questions do not have readily available external benchmarks. Alternatively, we can use quotas by socio-demographics as we have always done to create samples which look alike and "look like" the general population. But demographics are often not the most helpful or relevant stabilizers.

Take the absurd example where an online survey is conducted using only respondents who visit a website devoted to fishing. Since fishing is a reasonably popular sport among all age groups, we could produce a sample from this site that looked "right" in terms of demographics. But what sort of answers

(continued)

would they give? If the question is about brands of coffee bought, then their answers could well reflect the wider population—being interested in fishing doesn't affect the coffee you drink. But what if the questions were about media consumption? Anglers spend a lot of weekend time fishing, not watching TV. So the answers we get on TV viewing may not be anywhere close to the truth and you, the client, would never know this.

Using demographic quotas to ensure consistency works—as long as the stratification chosen is relevant to the topic of the questionnaire. And even when there is no correlation between, say, age and gender and perception of the product in question, the random nature of the sampling from a large population means that the chance of pulling a sample that has attitudes completely at odds with the last sample pulled is very small.

When using multiple online sources, however, you may be dealing with a number of relatively small universes that are not static and are possibly shrinking at each wave (especially if category resting is applied). Under these circumstances, it is much more likely that the true underlying constructs that go to define consumer choice in the category in question may differ in their prevalence in the sample pulled. It is not unusual for a product to be equally liked by both genders, across all age groups, regions and social classes—think Coke, for example.

Something, however, drives consumer preference for Coke over Pepsi. It is this "something," or a proxy for it, that needs to be controlled for or stratified on in the sampling process to ensure that two samples are truly comparable to one another, across time and across sources. Without it, the preference scores for Coke vs. Pepsi will fluctuate. In theory, the two samples balanced on the underlying "something" could be completely different in terms of age, gender, region and socio-economic factors and still come up with the "true" answer for Coke vs. Pepsi.

The two research studies reported in this paper are part of a research initiative Survey Sampling International (SSI) began about a year ago. SSI undertook the research to identify the components of this "something" and design a stabilization method to control the variation between sources which goes beyond socio-demographics. With an effective set of stabilization variables, identified at the individual participant level, it is not necessary to maintain exactly the same blend of sources from project to project. The number and proportion of sources can vary without causing variation in the results achieved.

If we can identify distinct psychographic, neurographic, and personality variables which are relevant to the way people answer surveys, we can control for them, and systematically broaden the diversity of our sample by finding sources which complement each other. For example it has been posited that people who enjoy participating in surveys from any source may have higher cognitive capabilities or a higher need for cognition (Cacioppo and Petty).

The first step in this process was to identify a broad set of factors which would define groups of people and potentially drive variance in behavior. We developed the questionnaire to include the major strands of thinking in relevant psychographic, neurographic, personality type, cognitive, learning and communication style. Elements included:

(continued)

- Personality traits, using traditional rather than Jungian style
- Music preferences (Rentfrow)
- Cognitive ability (Kahneman and Frederick)
- Geographic/personality alignment (Gosling and Rentfrow)
- Social Values (There are a number of models; Schwarz values were chosen because they were freely available and had some available international benchmarks, having been used successfully in Europe.)
- Need for cognition measures (Cacioppo et al)
- Neurographics measures
- Propensity factors such as the propensity to participate, risk averseness, attitude to privacy, propensity to share and information usage
- Chronotypes, i.e. “lark” or “owl” personalities (Roenneberg)
- Disruption/orienting reflex measure and habituation disruption. We considered these measures potentially insightful since they were not linguistically mediated. Examples used were background color changes and response time measurements (i.e. how much hesitation did a change in background color create?)

We chose each test as typical in class and in the absence of agreement from academics in the psychology of marketing as to the one definitive test in each discipline. We also chose them because, for some, external benchmarks exist, and they are, in general, less susceptible to cultural differences than other types of factors; any process we adopted needed to be international in scope and outlook. A number of additional measures were added. In all, 162 measures were tested.

Questions were created or adapted around each of the identified factors. These were tested against questions on topics such as technology adoption and usage, hobbies and interests, brand preference and loyalty, and reaction to advertisements—representing the most common types of general survey research that our clients are conducting today.

Sample came from the SSI US Survey Spot panel. As an aside, the questionnaires, at 34 minutes and 20 minutes, were longer than our ideal length. As we frequently preach about questionnaires being too long, we hesitated before placing them in field, especially as we offered no individual reward. However, the participant comments we received such as “Loved the variety of questions” and “The questions challenged me to look at myself,” and the fact that the drop rate on the second study at 7.6% was considerably lower than our average drop rate, bolstered our belief that there is a difference between “actual” and “perceived” survey length. If a survey is compelling and a topic for which people have engagement (talking about themselves for example!), they have greater tolerance for a longer survey.

Two thousand completes were obtained in the study, which produced two major findings. First, many variables explained considerably more variance in the dependent variables than socio-demographics alone. Here are some examples:

(continued)

		Explained variance with 16 best-performing independent variables	Explained variance with socio-demographics
When a new technology or product is invented that interests you, how soon after its release do you purchase?	7%	16%	2.3
How important to you are brand names when purchasing an item?	3%	8%	3.1
On how many of the past 5 weekdays have you read a newspaper?	10%	12%	1.2
Do you have a cell phone?	4%	7%	1.6
How frequently do you make or receive calls or texts on your cell phone?	13%	18%	1.4

Through iterative analysis techniques, each of the 162 variables we started with was tested on its power to "move the needle" on a variety of dependent variables. 16 variables were identified as a cluster which moved the needle most strongly.

Second, neither our cluster nor socio-demographics explained most of the variance. In other words, a person's presence in one or other of the clusters had little impact on whether they preferred the blue or the red shirt. Although it appeared that keeping this specific cluster balance constant would not provide gains which justified the additional complication of balancing on them every time we draw sample, this research study gave us intriguing hints about the potential of psychographic and neurographic measures to outperform socio-demographics as a sampling balancing mechanism.

The next phase of the research was, obviously, to re-test all this within a multi-sourced environment. In this phase, we benefitted greatly by working with Robert van Ossenbruggen, architect of the seminal 2006 NOPVO study. This study, which compared the performance of 19 panels in The Netherlands, was referenced in the design of the ARF Foundations of Quality study of 2009.

The sources used in the second study offered a diverse representation of the way people are brought into surveys and rewarded in the online environment:

- The SSI SurveySpot general population panel recruited online and rewarded with a variety of rewards, primarily sweepstakes and instant win prizes with some cash
- An SSI proprietary panel with a general population frame much of whom are recruited in person at shopping malls. Members have an affinity to their local mall and are rewarded with mall gift certificates.
- An SSI teen and young people panel, recruited to support gaming and entertainment research, and rewarded with prizes designed to appeal to a younger age group such as iTunes downloads

(continued)

- A database of people who are rewarded for performing various activities online
- A general opt in access panel
- A “river” source
- Another general opt in access panel

Results from study two confirmed and expanded on results arrived at in study one. Three major findings were determined.

1. Different sources, as expected, produced different results. The differences were found across a wide range of topics and questions:

Sample sources*:	A	B	C	D	E	F	SD	Variance
Universalism as principle	5.75	5.49	6.02	5.90	5.80	5.52	1.72	0.044
Prayer in school, agree	2.92	3.03	2.50	2.84	3.07	2.93	1.44	0.042
Hedonism as principle	5.48	5.24	5.55	5.54	5.52	5.09	1.88	0.037
Tradition as principle	5.62	5.62	5.98	5.79	5.72	5.41	1.69	0.037
Trust internet transactions	2.93	2.62	3.11	2.86	2.84	3.03	1.29	0.029
See self as disorganized, careless	2.36	2.32	2.53	2.59	2.39	2.74	1.57	0.026
Traditional role of women	3.41	3.56	3.10	3.37	3.37	3.50	1.26	0.025
See self as critical, quarrelsome	3.11	3.12	2.83	3.11	3.07	3.32	1.64	0.025

* One source was removed for clarity on this chart because it has an inherently different demographic make-up than the others.

We saw differences across the seven sources in many of the areas commonly studied in market research including attitude to new technology, use of organic produce, use of coupons, shopping decision drivers (whether buy on price or quality, financial habits and daily newspaper readership).

2. Balancing on education, income, employment and marital status, (the samples were pre-balanced on age and gender) doesn’t help us much in explaining the variance.

3. Using the psychological and neurographic variables does more to explain the variance, as we see in this summary chart for the entire study:

	No balancing	Balancing with socio-demographic variables	Balancing with psychographic, etc. variables
Between sample variation metric 1*	0.88	0.76	0.48
Improvement factor		1.15	1.83
Between sample variation metric 2**	18%	14%	9%
Improvement factor		1.29	1.92

*This metric was calculated as follows: 1. Means of IVs tabulated per source 2. Variance between sample means calculated 3. Sum of variances calculated.

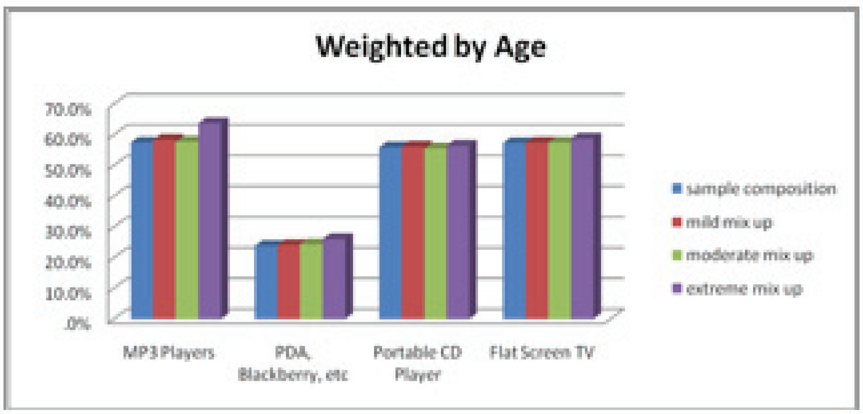
**This metric was calculated by calculating % of subsample means deviating 0.1 SD from total score.

(continued)

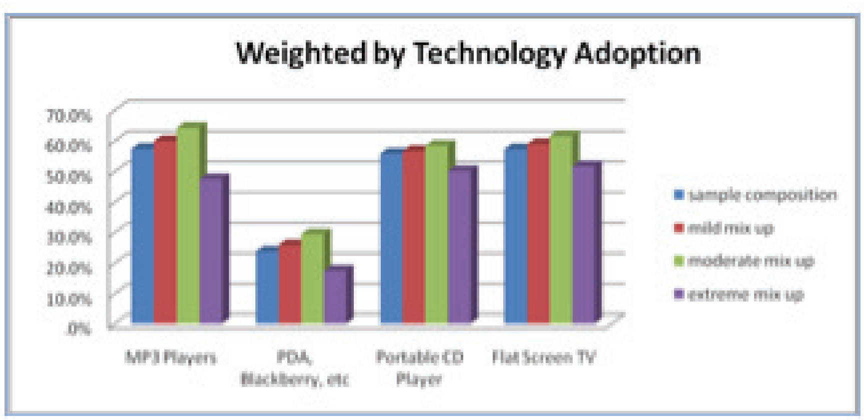
A Specific Example

"Willingness to try new things" was one of the variables with significantly different results across the seven sources. In looking at a simple cross tabulation of being willing to try new things and technology adoption, we see that people were less likely to adapt to new technology if they weren't open to trying new things (as we would expect). As we might also have predicted the willingness to try new things correlates with possession of technology such as a BlackBerry® cell phone or an MP3 player.

To look at the effect of different levels of demographic variables and compare the effect of different levels of psychographic variables, we created a set of samples with different mixes—from a "mild mix" to an "extreme mix." When we mix the levels of different ages within samples, the differences in ownership of various new technologies does not change as can be seen in the following graph.

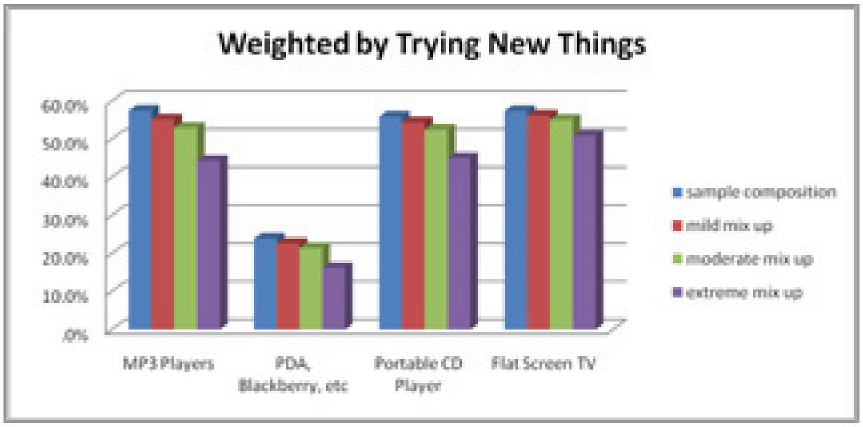


When we create a set of samples with different mixtures of levels of technology adoption, however, the resulting ownership of MP3 players, PDA, BlackBerry®, and Flat Screen TV varies significantly.



This is also the case when we create different samples with different mixtures of propensity to try new things.

(continued)



Again, this is important because the likelihood to try new things was one of the variables that was significantly different across many of the sources we tested.

Our advice from this example is to include a question about an individual's comfort level with technology when blending sample for research projects concerning technology.

This is one example of how we arrive at one element in a stabilization cluster. We expect to be using thousands of sources to create samples in the future, so we require a stabilization metric which will work at the individual person level, rather than the source level. We have used data from this and other tests to refine the group of variables identified in the first test and are currently testing the resultant calibration measure. We are creating a set of questions (question clusters) which can be asked of any potential research participant before they enter a survey. The clusters being developed will be customized for different research topics.

Viewing multiple sources of sample as a vast reservoir, the variable set we are developing will become the "state of the water" calibration measure. We use this measure when we dip our cup into the reservoir to test the water to monitor its health.

In Summary

1. Multi-sourcing is the future of online sampling; the world is changing and the traditional e-mail invitation access panel paradigm is no longer a fit.
2. Multi-sourced sample provides superior reach and diversity compared to online sample from fewer sources.
3. We can keep multi-sourced sample consistent by pre-profiling research participants.
4. Pre-profiling must be approached in a new way; the way the industry has pre-profiled in the past with standard socio-demographics doesn't do the job.
5. We have taken the first steps by creating a concise list of relevant variables that can be used for a broad range of research subjects.
6. Identification of bias and controlling for it in multi-sourced sample is a work in progress, but we see this approach as a big step in the right direction.

(continued)

Some Practical Steps for Blending Sample

1. Know your sources—understand how participants are recruited, where they come from, how they are rewarded, and means of communication.
2. Plan ahead and blend sample from the beginning of the project to avoid having to incorporate new sources midway through.
3. Include some calibration questions, such as the example of a willingness to try new things, customized for both the type and topic of your research study.
4. Discuss with your sample provider their techniques for blending sample. Ask if they use source smoothing as well as quality controls such as digital fingerprinting to ensure there is no duplication.

Finally, we are concerned with blending because we want to ensure that the opportunity to take surveys is placed in front of as large and diverse a population as possible. But the effort is in vain if our non-response and non-completion rates are so high that we introduce bias from those factors. We must also focus on making the questionnaire better for participants and ensuring the research experience is one that they will find, if not always thrilling, at least satisfying enough that they would be willing to do it again.

*This paper was first presented at the CASRO Panel Conference in February 2010.

References

Cacioppo, John T. and Richard E. Petty. "The Need for Cognition." *Journal of Personality and Social Psychology*, 42, 116-131. 1982.

Cacioppo, J. T., Petty, R. E., Feinstein, J, A., & Jarvis, W. B. G. "Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition." *Psychological Bulletin*, 119, 197—253. 1996.

Rentfrow, P. J., & Gosling, S. D. "The do-re-mi's of everyday life: The structure and personality correlates of music preferences." *Journal of Personality and Social Psychology*, 84, 1236-1256. 2003.

Peter J. Rentfrow, Samuel D. Gosling, and Jeff Potter. "A Theory of the Emergence, Persistence, and Expression of Geographic Variation in Psychological Characteristics." *Perspectives on Psychological Science 3* (5) Pages 339 - 369. 2008.

Fallig, Michael and Derek Allen. "An Examination of strategies for Panel Blending" Quirks July 2009.

Foundations of Quality Knowledge Briefs 1, 2 and 3 *Advertising Research Foundation*. www.thearf.org. 2009.

Kahneman, Daniel and Shane Frederick, "Frames and Brains, Elicitation and Control of Response Tendencies." *Trends in Cognitive Sciences*, Volume 11, number 2. Elsevier. 2006.

Schwarz, N., & G. L. Clore, "Feelings and phenomenal experiences." *Social Psychology: Handbook of basic principles* (2nd ed., pp. 385-407). New York: Guilford. 2007.

Till Roenneberg, Till, Anna Wirz-Justice and Martha Merrow, "Life between Clocks: Daily Temporal Patterns of Human Chronotypes." *Journal of Biological Rhythms*, Vol. 18, No. 1, 80-90. 2003.

More information about the NOPVO study can be found at http://www.nopvo.nl/english/english.htm

(continued)

(continued)