

Probability Samples vs. Volunteer Respondents in Internet Research: Defining Potential Effects on Data and Decision-Making in Marketing Applications

Vicki Pineau, Chief Statistical Officer

Daniel Slotwiner, Vice President, Panel Management
Knowledge Networks, Inc.

Abstract

As use of the Internet for commercial research grows, so too does the need for a deeper understanding of the medium's unique methodological challenges. One key area that has received little attention is sample source – the fact that most Web-based research utilizes “volunteer” groups rather than probability samples. While such groups can be used effectively for certain types of studies, treating them as equivalent to population-projectable samples can produce misunderstandings – and, potentially, mistaken business decisions. Pineau and Slotwiner review the potential sources of survey error in Internet research and provide case studies showing that, compared to conventional RDD samples, volunteer Web groups produce study outcomes that are substantially different – and would likely lead to different business decisions if not viewed in the proper context.

Contents

Introduction: Research Quality & the Growth of Web Surveys.....	1
Survey Quality, Survey Error and the Internet.....	2
Coverage Error.....	3
Self-Selection Error.....	5
Non-Response Error.....	6
Adapting the Web for Research, and Vice Versa.....	6
Case Studies: Demonstrating Bias Effects in a Market Research Setting.....	7
Alcohol Awareness and Usage Study.....	8
Facial Products Awareness and Behavior Study.....	13
Concluding Remarks and Future Research.....	15
References.....	15

Introduction: Research Quality & the Growth of Web Surveys

If the purpose of market research is to provide knowledge that yields smarter business decisions, then that knowledge must be accurate. But frequently, the details of how market research is designed and conducted are complex, discouraging those who simply want to know, “Can I trust these numbers?” In the face of this conflict, research quality often becomes a matter of perception; yet the factors that affect research quality continue to operate below the surface, shaping the information on which million-dollar business decisions are often made.

One of the fundamental elements of research quality is bias; bias is the difference between a statistic calculated in a given study and the true population value of that estimate in the target population. To minimize bias, researchers have traditionally attempted to create consumer samples that provide a reliable cross-section of a given population – a country, a county, a race, or a gender. Perhaps the most common and widely understood method for constructing such samples is Random Digit Dialing (RDD). RDD yields samples that are representative of the U.S. population (or some subset of the U.S.) because two conditions obtain: first, nearly all households have telephones (about 96%); and second, a comprehensive list of all area codes and three-digit exchanges exists. This allows researchers to draw random, or probability-based, samples and project results from that sample back to the entire population.

Research conducted solely on the Internet does not benefit from these conditions. The proportion of households in the U.S. with Internet access is about 57%, and they are different than the U.S. population as a whole (Couper 2000). Broader definitions of the Internet population yield only slightly higher estimates (see discussion below). This means that, at best, pure Internet samples can only represent that 57% of the U.S. population.¹ However, since there is no comprehensive listing of these households (or the email addresses of the household members), it is not possible to draw a random (probability)

¹ One could certainly argue that this definition of the Internet population is too narrow. We discuss our choice and other definitions below.

sample of them without resorting to traditional research methods. As a result, the majority of Internet research is conducted on pools of respondents who can be more accurately described as “volunteer groups” rather than a random sample.

The use of volunteer Internet samples has grown rapidly and is now widely accepted for a variety of applications. Government agencies, academic institutions, advertising agencies, consumer goods manufacturers, and media companies have all embraced Web research because of its speed, flexibility, video capability – and, frequently, its economical pricing. According to the January 2003 issue of *Inside Research*, overall spending for online research increased by 53% between 2000 and 2001, and by another 61% in 2002 – a level of growth that is expected to continue for several more years.

Consequently, market research is in a period of transition, as established methods and vehicles are applied to, combined with, and even replaced by Web-based work. Researchers who use non-representative (volunteer) groups of Internet respondents are usually aware of sample limitations; and for some studies, volunteers may be an acceptable alternative. Some also assert, however, that Web-only volunteer samples can be “balanced” so that they mimic RDD results; others have even gone so far as to suggest that data collected from volunteers can yield conclusions projectable to the U.S. population as a whole. These claims, viewed in light of the Web’s advantages as a survey medium, have convinced many that the differences between typical volunteer-based Internet research and established survey research methods are not substantial enough to merit concern, and that the two can be used almost interchangeably.

In this paper, we ask, “How comparable are the results between studies conducted on volunteers versus those from studies conducted on random samples?” The answer, we find, is that different methodologies do yield different results and would, by inference, yield different applications or decisions. To make judicious use of Internet research requires specific knowledge about how volunteer work differs from probability-based methods. In either case

complete knowledge of how the sample was constructed is a necessary pre-condition for analysis.

In the paper that follows, we will address each of these issues in turn. We will:

- review the factors that can introduce error into survey estimates;
- summarize the differences between the on- and off-line populations;
- evaluate some common techniques used to “approximate” representativeness when using volunteer Internet results; and,
- present case studies demonstrating that sample source often has a tangible and important impact on study outcomes – and business decisions – in a market research setting.

Survey Quality, Survey Error and the Internet

Survey quality is directly related to the level and effects of survey errors. The greatest challenge faced by researchers is to simultaneously minimize all types of survey error, so that the study results are as accurate as possible, can be duplicated, and meet the survey measurement objectives.

Bias is introduced in survey estimates to the extent that those not covered, not recruited, and/or not surveyed are *different* from those who are covered, are recruited and do respond. In the world of Internet research panels, sample bias typically comes from any of three main sources:

- **Coverage Error:** Bias introduced into a study when certain groups are excluded (e.g., excluding non-Internet households, or non-telephone households).
- **Self-selection Error:** Bias that occurs when a panel is made up of volunteers (“self-selected”) instead of being derived from a designated sample.
- **Survey Nonresponse Error:** Bias due to nonresponse to a specific survey experience.

All surveys (including probability-based studies) have differing levels of bias. The end goal is to minimize the level of each error, measure any potential bias, and adjust for that bias to the extent possible. Summarized in the table below are approximate measures of the success in minimizing survey error in volunteer Internet estimates as compared to RDD telephone research and the Web-based Knowledge Networks Panel, which is derived from an RDD sample of the full U.S. population. In the sections below, we describe the potential impact of these types of errors on survey results.

Table 1: Summary of the Levels of Success in Coverage, Panel Recruitment, Survey Response

	Coverage of U.S. Population	Panel Recruitment	Survey Response
KN Panel	96% ^a	37%	50-90% ^e
Volunteer Internet	57% ^{b,c}	.02% ^d	15-35% ^e
RDD Telephone	96%	N/A	25-50%

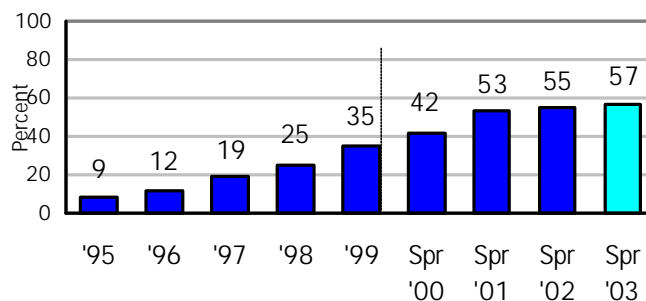
^a Non-telephone HH not covered.
^b Non-Internet HH not covered.
^c This may be overly generous, as the real coverage could be defined as “those who see the recruitment impressions.” This will always be just a portion of all those on-line.
^d Although it is a critical measure of data quality, those who maintain volunteer Internet panels do not provide information about the success of their recruitment strategies. The best data available come from Alvarez, Sherman and VanBeselaere 2003, p.31 who show click-through rates of around .02%.
^e 7- to 30-day fielding period

Coverage Error

Coverage error occurs when some persons are omitted from the list or frame used to identify members of the study population. Although the Internet has become more prevalent over the past decade, significant differences persist between the on- and off-line populations, with a large portion of the American population remaining “off-line.” A pre-condition to measuring these differences, however, is defining what it means to be on- versus off-line. For the purposes of this paper, we follow the U.S. Census and

define the on-line population as “those with access to the Internet from home.” While this is the narrowest definition, it is also the most concrete, stable and easy to measure metric. Some alternative definitions are “those with an email address,” “those who have been on-line in the past month” or those “who have access to the Internet from home, school or work.” Depending on the research question and context, any of these definitions may make sense. The crucial point, for this paper, is that the broadest definitions of Internet diffusion suggest that 30% to 40% of the U.S. population remains off-line. Chart 1 and Chart 3 (see below) summarize some of these estimates.

Chart 1: Trend in On-Line Use by Total Households, 1995 – 2003



Base: Total households

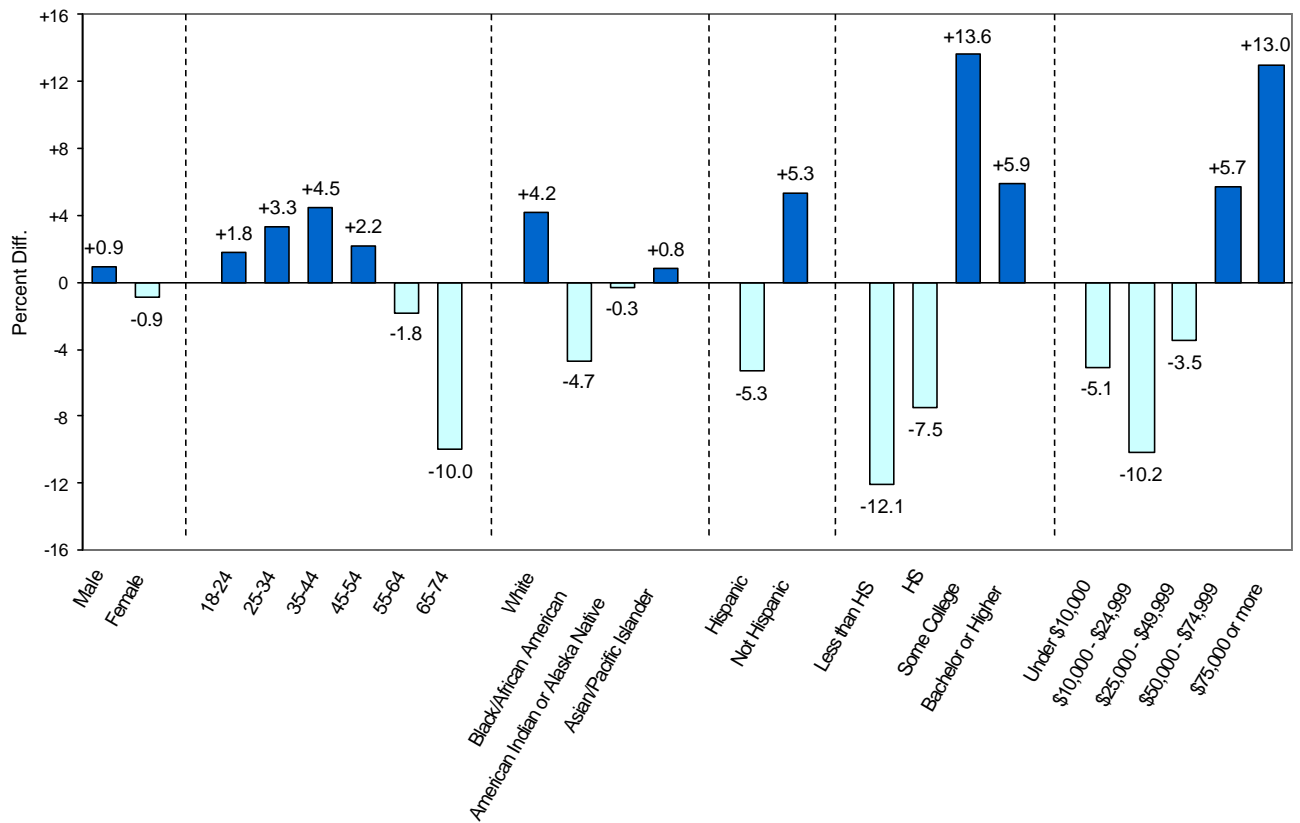
The data reported in Chart 1 come from Knowledge Networks’ *The Home Technology Monitor*TM, a continuous tracking study conducted via telephone of 3,000 households per year. In addition to providing estimates of Internet adoption, Charts 1 illustrates the important point that Internet penetration has slowed in recent years – a conclusion that holds no matter what definition of “on-line” is used.

Chart 2 (top of next page) presents data from the U.S. Census Bureau’s Current Population Survey² and summarizes some of the demographic differences that still exist between the on- and off-line populations.³

The most significant differences are among older Americans (including baby-boomers), African-

² <http://www.bls.census.gov/cps/computer/sdata.htm>
³ “On-line” households are those that have a computer in the household that can connect to the Internet.

Chart 2: Percentage Difference Between U.S. Population and U.S. Online Population (Households with PC's Connected to the Internet)



Note: Positive numbers indicate over-representation in on-line population.

Americans, Hispanics, those with incomes below \$50,000 and those with lower levels of education. Because these groups are less likely to be on-line, randomly selected Internet-only samples will under-represent them relative to the U.S. population as a whole. More significant than the demographic differences between the two populations, though, are the behavioral, attitudinal, and other, unknown differences that exist. Knowledge Networks research has shown, for example, that Internet populations have significantly different opinions on topics such as abortion and political affiliation,⁴ and that their use of

media and ownership of home media technology diverges from non-Web homes.

Another source of valuable probability-based data about consumers and the Web is the Pew Charitable Trust, which conducts ongoing RDD telephone research (n ~ 15,000) for its Internet and American Life Project. In a Pew report from August 2003, Spooner *et al.* (2003, viii) note that the Internet population is developing in ways that distinguish it from the U.S. population as a whole. The authors summarize their findings as follows:

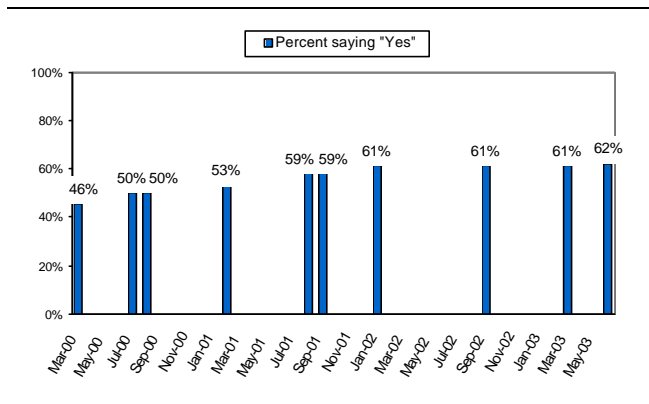
- Internet penetration continues to be unevenly distributed in different parts of the country.
- Regional variations in Internet use reflect differences in education and income levels.
- The race, age, and gender of Internet users also exhibit some distinct regional variations.

⁴ Among the Internet population, 21.4% strongly agree that abortion is murder as compared to 25.6% among the NonInternet population. 40.2% of the Internet population considered themselves to be Moderate as compared to 46.7% among the NonInternet population.

- Some online activities such as email are universally popular, but others, including online shopping, are favored only in certain regions.
- Experienced Internet users throughout the country log on more often during the day but limit their time online.

The Pew data, like the data from Knowledge Networks' *The Home Technology Monitor*TM, also indicate that the rate of Internet adoption has slowed dramatically. This is particularly interesting in that they employ a much broader definition of "being on-line". Their measure of being on-line classifies anyone who has ever been on-line to access the Internet or send and receive email as being on-line. In March of 2003, they estimated the on-line population to be about 61% of the U.S. (see Chart 3). Note that Chart 1 also confirms that the rate of growth of the on-line population has dropped dramatically. Again, quoting a Pew report (Lenhart *et al.* 2003, 3): "It may be that we have reached a point where the adoption curve has peaked and the market is no longer working to bring online new groups of Internet users."

Chart 3: Do you ever go online to access the Internet or World Wide Web or to send and receive email?



The difference between the population surveyed and the population not surveyed can have real consequences if one attempts to generalize *beyond* the surveyed group, *even if the sample sizes are in the millions*. The well-known example of the Literary Digest political poll conducted in 1936

illustrates this point well. As summarized by Don Dillman (2000, 335) and others, the Literary Digest attempted to predict the winner of the 1936 Presidential election by mailing sample ballots to a sample drawn from telephone directories and auto registration records. In response to this mailing, the Literary Digest received 2,266,566 returned ballots (about a 23% response rate). Of these ballots, 1,293,669 (57%) were for Alfred Landon and the remaining 972,897 (43%) were for Franklin Roosevelt – leading the Literary Digest to predict that Alfred Landon would win the election.

Of course, Franklin Roosevelt went on to win the election by a landslide, with more than 60% of the popular vote – leaving the editors of the Literary Digest to wonder where they went wrong. By attempting to generalize the results of their survey to the entire electorate, the magazine editors erroneously assumed that those included in their sample frame (households with telephones) were not substantially different than those who were excluded (everyone without telephones). In fact, we know today that those with telephones were wealthier and more likely to be Republicans. In this instance, low coverage of the population of interest led to an inaccurate prediction of the outcome.

As with the telephone in its early days, the Internet now reaches only a select portion of the population – one that is clearly different, but whose attributes cannot be measured or balanced. Thus, any Web-based attempt to recruit respondents will create inherent, unpredictable anomalies in the usability of the data. Coverage error is inherent in non-RDD Web research.

Self-Selection Error

In order to be representative of the population, or some segment thereof, research needs to be conducted among a designated *sample*. The sample is chosen based on a thorough knowledge of the study's universe, or coverage area. In population-projectable research using a panel, one sample needs to be drawn when recruiting the panel, and a second for each individual study. The "success" of the research will

depend in part on the ability of the researcher to obtain cooperation at both stages of recruitment.

At present, the population of the Internet is statistically indefinable; unlike telephone numbers, Internet connections and/or email addresses do not provide a method for uniquely representing all of those on the Web. And e-mail addresses are not created according to a standardized system, as phone numbers are. Without this baseline structure, it becomes impossible to create a probability-based sample that represents the Web population in a statistically acceptable way.

Recruitment for nearly all Internet panels and one-time surveys is conducted on a “volunteer” basis. Advertisements guide anyone who may see them to a signup site on the Web, and others can go directly to the survey site after hearing about the opportunity from family, friends, co-workers, the media, or other sources. In fact, many on-line panels reward members for “referrals”. Volunteers are typically not screened out of panels; some screening may be done for specific surveys, but that does not create a unique and probability-based sample. In addition, it allows for respondents to self-select into surveys and panels that deal with topics of interest to them.

Statisticians refer to these types of samples as “convenience samples,” and their defining characteristic is that their universe characteristics are unknown. Studies conducted with these types of samples are not appropriate for making projections and should not be used when the stakes are high. They can, however, play a useful role in the “idea generation” phase of research. In some instances, convenience samples may be the only feasible way to get any information at all (for example, when incidence is extremely low). In these cases it is still imperative that conclusions remain confined to the sample at hand and all aspects of how the sample was selected get reported in detail alongside the results.

Non-Response Error

Non-response – failure of a selected unit to participate – can occur in reference to panel membership, a specific panel study, or a survey among one-time

respondents. *Response rate* is just as important for Web research as for telephone, but it is often overlooked or dismissed in the Internet world. Non-response affects study outcomes by the degree to which respondents and non-respondents differ, as well as the relative size of the non-respondent pool. So, even if substantial differences exist between non-respondents and respondents, if the level of nonresponse is low (10% or less), the impact on final outcomes is minimal. This generally defines the researcher’s goal – minimize nonresponse in order to minimize potential bias in survey results.

The issue becomes more complex if research vendors do not take steps to maximize – through “refusal conversions” – the cooperation rate of those invited to participate. For example, recruitment conducted via banner ads and pop-up windows is rarely followed up with conversion attempts; thus, the characteristics of those invited to join the panel are unknown, and the response rates for these invitations often fall in the 1% to 2% range or below. The question researchers must ask is whether the other 98% to 99% of people who do not join are different than those who do.

Recent research into the effects of non-response error has found little evidence that increased levels of non-response have substantive effects on study outcomes. However, these studies did not investigate cases in which non-response reached levels of over 80% (see Curtain, Presser and Singer 2000). Similarly, these studies did not investigate the impact of conducting surveys with absolutely no follow-ups to the first contact. Clearly, the levels of non-response experienced with Internet research merit closer investigation. In order to meet this challenge, researchers must adopt guidelines for calculating and reporting measures of non-response so that “apples to apples” comparisons can be made across surveys.

Adapting the Web for Research, and Vice Versa

Clearly, the Internet offers advantages that make it a natural choice for research applications; and “volunteer” Web research can serve as a broad guide

or a source of ideas in applications where precision is not the first priority. As this paper has shown, however, non-RDD Internet studies pose a substantial challenge when it comes to delivering samples that can meet standards of quality for decision-makers government, academic, and commercial work. To allow the basic data analysis and weighting that can only be achieved with representative samples, Internet researchers have a limited number of choices:

1. Using a multi-mode approach and a probability-based sample where the first contact is made by phone, mail or in-person interview. The e-mail address is obtained from as many respondents as possible for Internet survey administration. Persons without Internet access must be interviewed using an alternative mode – an inordinately lengthy and costly procedure.
2. Recruiting an RDD sample via telephone, and providing Internet access to those in the sample who do not already have it.

The latter methodology is currently in use by Knowledge Networks, which maintains the only projectable Internet panel in the U.S. Using an RDD sample of the U.S. population, Knowledge Networks recruits potential panelists by telephone, and those who do not have the Internet are provided with an MSN-TV device and free Web access. In this way, the KN Panel represents the full spectrum of U.S. consumers, and actual response rates – as well as sample probabilities and reliability estimates – can be calculated. The Knowledge Networks method eliminates non-Internet coverage bias and allows researchers to accurately gauge the potential for self-selection and non-response bias.

Some researchers and sample vendors assert that the differences between Internet and non-Web RDD samples can be fixed after the fact through quota sampling, weighting, and balancing. These techniques are based on demographic approximations of the differences between on- and off-line populations; by

adjusting Web results to create a respondent group more like the non-Web world, this theory argues, the effects of bias can be minimized.

But the bias inherent in Web-only volunteer samples affects their results in unpredictable ways, no matter how carefully respondents are quota sampled or how rigorously balancing is applied. Balancing a non-representative sample using ratio adjustments to demographic, geographic, and attitudinal data from other sources is only guaranteed to force the specific estimates that are used in the balancing to look similar; it does not ensure that the key estimates of interest collected in the survey are representative. Balancing a probability-based sample to reliable benchmarks is an oft-used and statistically valid technique (Deming 1943); however, balancing a non-probability sample can yield erratic results.

Similarly, while it may seem logical to assume that larger samples or respondent groups will compensate for a lack of representativeness, the opposite may actually be the case. Problems with sample bias maintain, regardless of the sample size, and may even be magnified where larger respondent groups are used. As the Literary Digest poll (n=2,266,566) made explicit, obtaining responses from a large number of respondents is no substitute for good coverage.

Case Studies: Demonstrating Bias Effects in a Market Research Setting

To demonstrate the impact that coverage, self-selection and non-response error can have on market research, we have conducted several case studies. The goal of these case studies was, simply, to understand the extent to which results from a typical study would (or would not) differ when that same study was conducted on an RDD-based sample versus a non-random Internet sample – and, by extension, what impact those differences might have on business decisions.

Alcohol Awareness and Usage Study

The first case study was conducted on behalf of a well-known alcohol manufacturer that uses Knowledge Networks panel studies as benchmarks for some of its market research. This client commissioned Knowledge Networks to conduct an incidence study focusing on consumption of alcohol at home and away from home in order to assess market opportunities. A large sample was fielded on the Knowledge Networks probability-based panel in order to obtain “readable” sample sizes for lower incidence alcoholic brands.

Two volunteer e-mail lists were also administered virtually the same questionnaire as the Knowledge Networks Panel sample. These lists came from well-known suppliers in the marketing research industry, and from companies that typically supply other research companies with sample. For simplicity, we will refer to the volunteer data as EM-1 and EM-2. Table 2 summarizes the total sample sizes and qualification rates (which differed greatly between the KN sample and EM-1 but could not be calculated for EM-2). All three samples were exclusively of males in the 21 to 27 age range.

Table 2: Study Overview

	KN Sample	EM-1	EM-2
Total Completed	541	318	Unknown
Total Qualified	202	219	168
Qualification Rate	37%	69%	Unknown

In order to assess the comparability of the three data sources and to demonstrate the extent to which these various methodologies lead to *different*

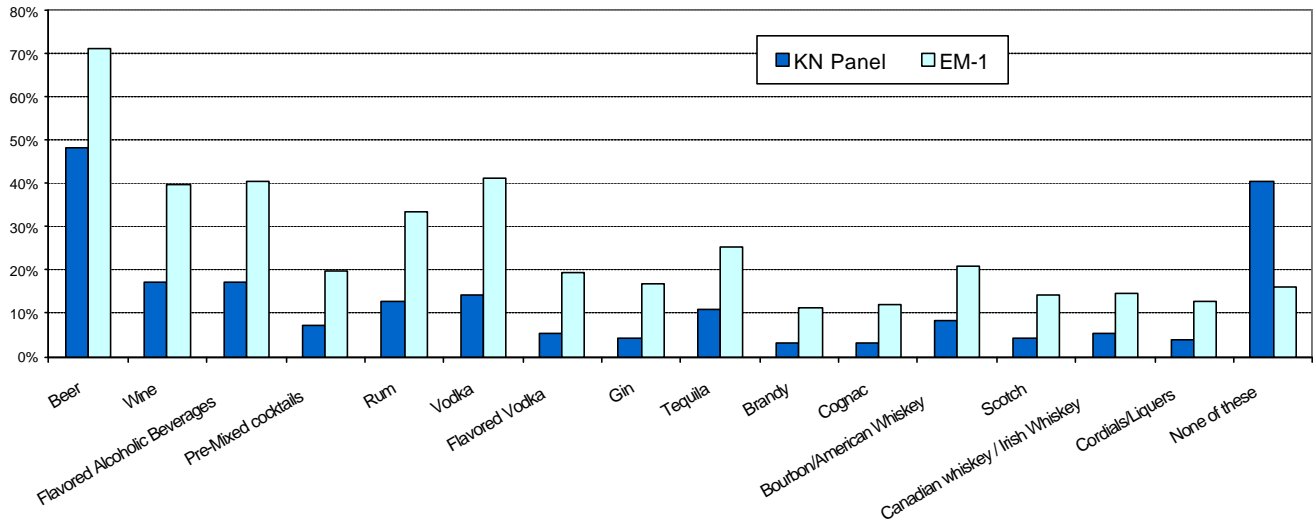
measures of incidence, we undertook three analyses. First, we examined the data to see whether or not the different vendors produced comparable levels of on- and off-premise alcohol consumption. Second, we evaluated whether the data exhibited the same relative distributions across the vendors. Finally, we investigated the extent to which weighting the EM-1 and EM-2 samples changed the results of the first two analyses. In all cases, we found that the data from EM-1 and EM-2 differed significantly from the KN data. We found that the basic results – that both Internet volunteer samples yielded consistently higher levels of usage and awareness than the KN sample – are manifested in all the data collected. These results are consistent with the traditional notions that strictly volunteer samples bring the participation bias of its volunteers.

Comparability across Samples

Chart 4 (top of next page) summarizes the estimates derived from the first question in our survey – Which of the beverages listed below have you consumed in your own home or someone else’s home in the past month? Data are presented for the KN sample and from one of the e-mail list vendors.

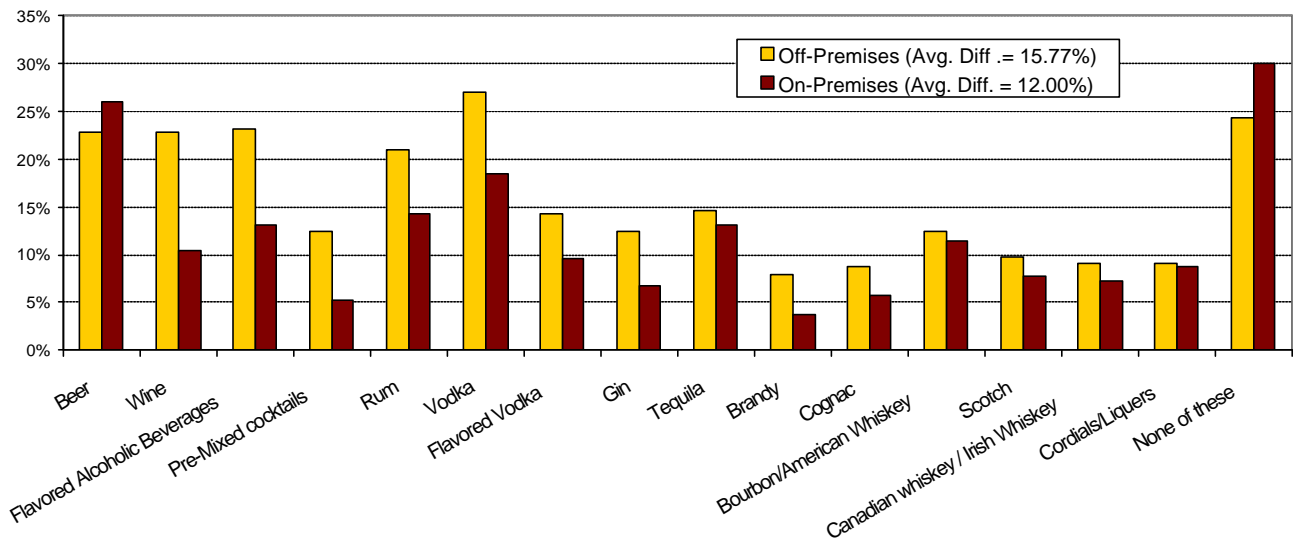
The average difference between the KN estimates and the estimates from EM-1 is about 16%. Reproducing this chart for the second question – “Which of the beverages listed below have you consumed outside the home, for example in bars, clubs or restaurants in the past month?” – yields nearly identical results (the mean difference in the estimates was 12%). Chart 5 summarizes these differences by category for the measures of on- and off-premises alcohol consumption.

Chart 4: Off-Premises Consumption



Note: We could not use the data from the second source for this chart because that vendor would only provide data for the qualified completes (hard liquor drinkers), and this table is based on all completes.

Chart 5: Absolute Differences

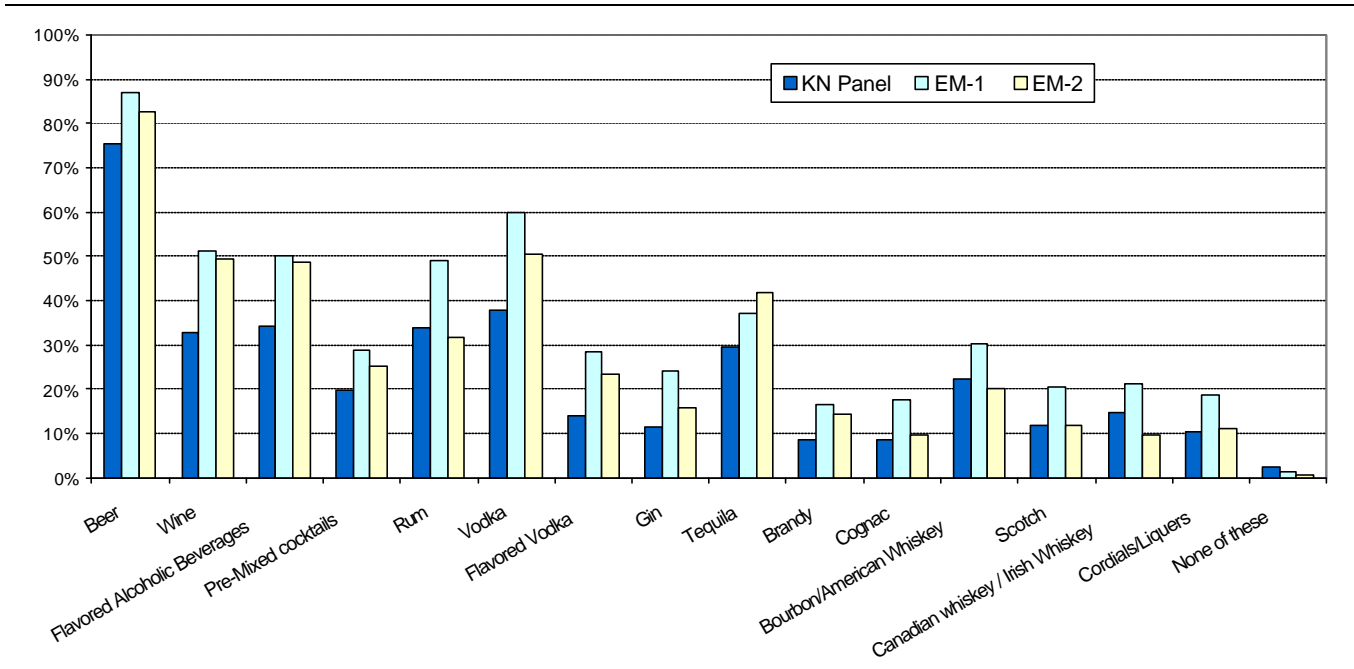


The measures of on- and off-premises consumption, then, show clear differences across the two samples. EM-1 estimates are higher for alcohol consumption than the KN panel estimates. The pattern holds true for practically every type of alcoholic beverage examined in the study (beer, wine, flavored alcoholic beverages, pre-mixed cocktails,

rum, vodka, flavored vodka, gin, tequila, brandy, cognac, bourbon, scotch, Canadian whiskey, cordials).

An examination of just the qualified completes yields similar results. Here the differences are seen across all three vendors. Chart 6 (next page) is identical to Chart 4 except that the base is qualified completes (allowing us to include the EM-2 data).

Chart 6: Off-Premises (Qualified Completes)



Calculating the mean absolute difference in the estimates, based on the qualified completes only, shows that difference between the KN data and EM-1 is slightly less (rather than 16%, it is 11%). The average difference between the KN sample and EM-2, though, is less – around 6%. Although the divergence in estimates is minimized when one restricts the analysis to qualified completes, it does not disappear, and the basic findings – that EM-1 and EM-2, on balance, provide higher levels of off-premises consumption than the KN data – hold.

It is interesting to note that significant differences exist between EM-1 and EM-2 despite similar methodologies. This was unexpected and raises a question (beyond the scope of this analysis) about the extent to which EM-1 and EM-2 are based on methodologies that yield reproducible results.

Although we have presented here only a limited amount of data from this study, we have analyzed all data and find that the results presented are typical of those that would be found if one proceeded question by question through the entire survey.⁵

⁵ Anyone wishing to pursue this claim should feel free to contact the authors.

Relative Distribution of Estimates

In the analysis above, we simply compared levels of on- and off-premises consumption across the various vendors. It could be argued that these differences, however great, are not of primary importance to the marketing issues at hand. The logic here is that the absolute values are of little consequence to the ultimate business decision, and that what actually matters is *relative differences* (i.e., “within sample variance”). In this case, for example, the argument would be that what’s important is not whether 71% versus 48% of people consumed beer at home in the past month, but rather what the difference is between wine and beer consumption at home during the past month. Both the KN sample and EM-1 yield comparable numbers (31%) for this measure. That is, as Table 3 shows, both the KN and EM-1 data indicate that beer consumption at home exceeds wine consumption at home by a margin of about 31%.

Unfortunately, this consistency between the two data sources does not extend beyond the comparison of beer to wine. When comparing beer to pre-mixed cocktails, for instance, the KN sample shows that beer consumption exceeds that of pre-

mixed cocktails by nearly 41%. However, the EM-1 sample indicates that beer consumption outpaces pre-mixed cocktail consumption by 51%. Thus, there is a 10 percentage point difference between the two estimates of the same statistic. In fact, the estimates from the two sources (for off premises consumption shown in Chart 4) are only correlated at .7, suggesting that they vary in concert but not perfectly. This fact is confirmed by the table below. When one compares beer consumption to each of the other categories for the two samples and takes an average of the differences, the mean difference is still a significant 9%.

Table 3: Relative Difference Between Beer Consumption and Other Categories by Vendor

	Beer v. Others (KN)	Beer v. Others (EM-1)	Absolute Difference
Beer	0.0%	0.0%	0.0%
Wine	31.1%	31.2%	0.1%
Flavored Alcoholic Beverages	31.1%	30.8%	0.3%
Pre-Mixed cocktails	40.9%	51.3%	10.4%
Rum	35.6%	37.5%	1.9
Vodka	34.1%	29.9%	4.2%
Flavored Vodka	43.0%	51.6%	8.6%
Gin	43.9%	54.4%	10.5%
Tequila	37.3%	45.6%	8.3%
Brandy	44.9%	59.8%	14.9%
Cognac	45.0%	59.2%	14.2%
Bourbon/American Whiskey	39.9%	50.3%	10.4%
Scotch	43.8%	56.9%	13.1%
Canadian Whiskey / Irish Whiskey	42.8%	56.6%	13.8%
Cordials/Liqueurs	44.4%	58.2%	13.8%
Mean	39.8%	48.1%	8.9%

Table 3 can be reproduced for qualified completes only. When this analysis is restricted to just the qualified completes, the estimates of three vendors correlate at .9, but there are still perceptible differences in the relative distributions of these estimates. The mean difference between estimates from the KN data and EM-1 drops to 5%. The mean

absolute difference between KN data and EM-2 is slightly higher at 6%.

Sample Balancing

Market researchers often use ratio adjustments to “balance” the data obtained from non-probabilistic methods to independent demographic benchmarks, such as the U.S. Census. Since this is a common practice, and the differences between the KN sample, EM-1 and EM-2 were so significant, we reproduced this procedure in order to understand the impact it would have on the comparison.

Using U.S. Census data, we weighted the EM-1 data so that the weighted frequencies for race, age, education and geographic region matched those of the U.S. Census. We then reproduced Charts 4 and 5 (as Charts 7 and 8) from above using the weight. The findings summarized above changed negligibly. Significant differences were still evident between the KN panel sample and EM-1 data.

The average differences between the two datasets for on- and off-premises usage (12% and 15%, respectively) were nearly identical to their unweighted counterparts (12% and 16%). As can be seen on Charts 7 and 8 on the next page, some estimates even moved further apart. For example, the unweighted estimate for past-month, off-premises consumption of beer from EM-1 was 71% (the KN panel estimate was 48%). After applying the weights, the EM-1 estimate changed to 75%.

Conclusions & Discussion

All methods of research contain the potential for bias that can skew estimates and lead to erroneous conclusions and measures. It is the job of the researcher to minimize the potential sources for bias and to make their presence known to the extent possible. In the data collected for this alcohol incidence study, it is clear that the different methodologies produced different incidence estimates.

Incidence studies like this one are conducted to measure the brand/category usage in any particular category (alcohol, make-up, frozen foods, etc.). They

Chart 7: Off-Premises (EM-1 Weighted)

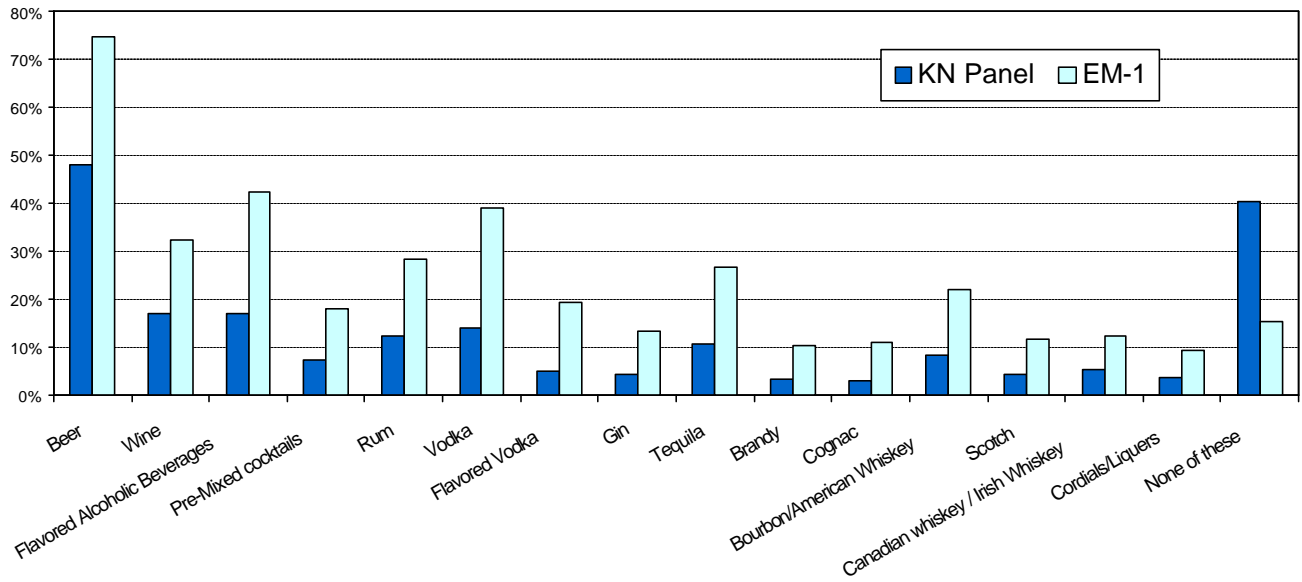
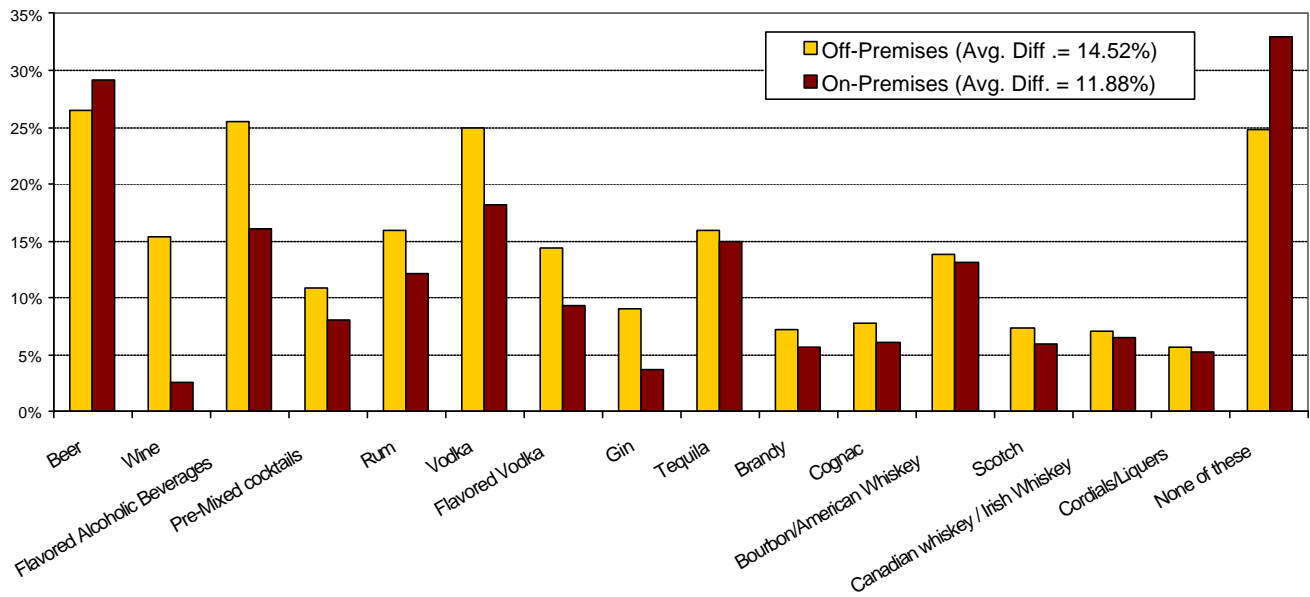


Chart 8: Absolute Differences (EM-1 Weighted)



provide snapshots of the competitive landscape and expand the understanding of brand managers well beyond what can be inferred from their own purchasing data. This overview of the category informs various decisions and serves as the basis for conducting further research. In addition to being useful for identifying market opportunities and estimating market size, these studies are the basis for

calculating the costs of doing in-depth research on sub-sets of category users. Needless to say, these subsequent studies often inform the strategic decision-making process and ultimately play a role in the allocation of advertising dollars, R&D investments and go/no-go decisions of all varieties. Thus, for a study of this type, where a premium is placed on the projectability, accuracy and reproducibility of the

findings, this alcohol manufacturer found it prudent to use a probability-based sample.

The results from this study show that the impact of the different methodologies is significant. If, for example, this manufacturer were trying to estimate the overall size of the market for flavored alcoholic beverages and intended to invest heavily in product development if the size of the market exceeded 25% of males age 21 to 27, the final business decision would have been different depending on whether they relied on KN panel data or EM-1. Using just KN data, the manufacturer would have estimated the market size to include just over 15% (see Chart 4) of the sampled group and would have directed their investment elsewhere. If the manufacturer relied on the EM-1 data, they would have estimated the market size to be over 40% of the target population and would have made the investment. Although this is a simplified example of how these data are used, the point is not weakened – data collected through awareness and usage studies inform strategic decisions in a very direct way.

Another way in which awareness and usage data are used is to further the research and market-intelligence-gathering process. In this sense, the data from these studies are used to allocate research dollars for more detailed and specific studies. Often these follow-up studies are conducted on the basis of incidence rates and other measures from the incidence study. Thus, there is a multiplicative effect to the extent that prior research informs subsequent research. For example, consider the qualification rates observed in the KN data versus EM-1 (37% and 69% respectively). If a marketer wished to conduct research solely on hard liquor drinkers, he or she would estimate wildly different field costs based on an incidence rate of 37% versus 69%, causing them to either over or underestimate the cost of the research. If timing were important to the study in question, an over-estimation of incidence could result in under-estimating the field time required and could jeopardize the success of the entire project and the chances for addressing the business challenges that necessitated it.

Facial Products Awareness and Behavior Study

The second side-by-side study conducted to compare the data from a non-probability-based Internet methodology with that from a probability-based sample was a facial products awareness and behavior study. This study was conducted on females age 13-69 who used certain types of facial products; the incidence was estimated to be about 2% in the population.

A low-incidence study of this nature offers several opportunities for research. First, these studies are often quite costly to conduct on probability samples; therefore, convenience sampling (such as mall intercept methodology) is often favored. The question of whether or not probability sampling still yields different results at low incidences is an important one. Another reason for choosing a low-incidence study was to examine the performance of the two samples in an increasingly popular type of research – an awareness and usage study of a newly introduced product. Because these surveys target early adopters, they are by definition almost always low-incidence studies. Internet research provides marketers with cost-effective access to low-incidence groups and is offering marketers insights at the earliest stages of a product launch. For this reason, it is one of the fastest-growing areas of Internet research; but it is also a relatively new form of research that requires serious study.

A well-known manufacturer of facial products (such as cleansers and moisturizers) retained Knowledge Networks to conduct a detailed study of a newly introduced product. The objectives of this research were to:

- Determine the penetration of the new line in the U.S. marketplace 6 months after launch
- Determine the effectiveness of marketing and communications efforts in generating awareness and initial usage of the product line
- Profile early triers of the product line to refine targeting efforts

- Assess product performance and satisfaction among early users
- Evaluate the line on key end-benefits delivered to early users

These data were intended to supply answers to important questions such as: Who are these early triers in terms of demographics and behaviors? Does the profile of early users match initial marketing efforts for the line, or does the targeting need to be refined? To answer these questions, it is critical to obtain accurate measures of incidence and reliable data on the demographic profiles of users. If the target group is not, ultimately, representative of all new users, then measures of product performance and benefits may be skewed as well.

Results

The KN Panel data indicate that U.S. marketplace penetration is about 1.2%; that is, 1.2% of those surveyed said they used the product one or more times per week. This figure is appropriate for a product six months post-launch in a highly fragmented category such as HBA cleansers/soaps/moisturizers/anti-age creams. EM-1, however, yielded substantially different measures of brand usage; the line penetration number was 3.3% -- nearly two and a half times the KN data estimate.

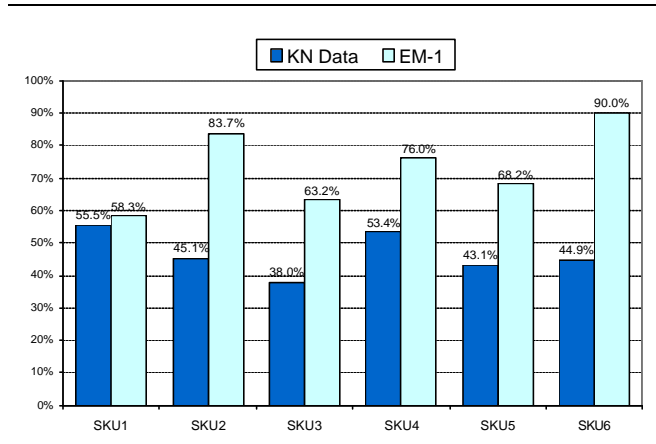
Other differences were clearly evident between the two samples. For example, the profiles of brand users described by the two sources of data show notable deviations. Like Internet users in general, early brand triers identified via EM-1 appear to be younger, more educated, wealthier, and tend to have more children than their KN data counterparts; they are also less likely to be African American or Hispanic.

From a behavioral perspective, data gathered by KN regarding early triers of the newly introduced brand also showed distinct differences from that collected through EM-1. Relative to EM-1, the KN data showed brand users to be less-frequent users of firming or anti-sagging products in general, less likely to have ever used specific SKUs in the category, less

likely to be daytime users of the individual SKUs, and less likely to have personally purchased the product themselves.

All of the behavioral differences noted above undoubtedly contributed to the subsequent differences in the evaluations of the product. As Chart 9 makes clear, product satisfaction – a key evaluative metric and a critical driver of repeat purchase – is significantly higher among those surveyed in the EM-1 data than among those from the KN data. The mean absolute difference in the six evaluations shown in table six is 26.6%. Those in the EM-1 data are also more likely than those in the KN data to believe the line is more unique than other cleansers.

Chart 9: Individual SKU's versus Broader Category – Percent Saying New Brand is "Better than others"



Conclusions & Discussion

The differences observed in this case study have profound implications for the business decisions at hand. For instance, the disparity between the estimates of brand usage could lead the manufacturer to draw dramatically different conclusion about the success of their product. On the one hand, using just the EM-1 data would lead them to the conclusion that the product is performing dramatically better than the average product performs 6 months after introduction. As these data are used to make explicit decisions regarding advertising and marketing efforts, such a conclusion could lead to decreased support for the brand when, in fact, its performance is average for the

category. If, in fact, the truth was closer to that measured by the KN data, the decision to reduce support for the brand could have serious consequences.

The extreme difference in the profiles of early adopters painted by the two data sources could also yield divergent business decisions. For instance, understanding the basic characteristics of early adopters is key to assessing and refining marketing strategy. It affects all decisions related to the purchasing of targeted media for marketing and communications efforts. As a result, misunderstanding this group could lead to an unwarranted change in course or lack thereof.

Concluding Remarks and Future Research

As the popularity and acceptance of on-line research grows, the importance of understanding its unique characteristics, advantages, and drawbacks grows with it. This paper is an initial attempt to address the sampling and selection component of this new methodology, and we have only scratched the surface of the issues and implications at hand. Ultimately, we believe that no form of data collection is right for every circumstance and that, when used cautiously, there is a place for non-probability-based Internet samples; the question is, under what circumstances and with what tradeoffs.

To this end, this paper is the beginning of what will be a sustained research program that will focus on the differences between probability-based Internet research and non-probability Internet research. The research will focus on setting guidelines for when it is appropriate to use non-probability data and when it is not. Furthermore, it will address procedures and methodologies that can be used to combine the two types of data in a manner that allows probability-based data to inform non-probability-based data (see Rivers, Pineau & Slotwiner ASA 2003).

The first step in extending the research conducted here will be to compare the results from studies conducted on various methodologies to known

population values. These “natural experiments” will allow us to compare estimates to known quantities. Elections offer a rare opportunity for such studies, but we will focus on areas with more direct relevance to marketers. In the meantime, we hope that this paper stimulates further research and interrogation of Internet methodologies and encourages all researchers using the Internet for data collection to refine their conclusions based on the limitations of the data and to use extreme caution when projecting to the U.S. population or subpopulations.

References

- Berrens, R. P., A. K. Bohara, et al. (2003). “The Advent of Internet Surveys for Political Research: A Comparison of Telephone and Internet Samples.” *Political Analysis* 11(1): 1-22.
- Couper, M. P. (2000). “Web Surveys: A Review of Issues and Approaches.” *The Public Opinion Quarterly* 64(4): 464-94.
- Couper, M. P., M. W. Traugott, et al. (2001). “Web Survey Design and Administration.” *The Public Opinion Quarterly* 65(2): 230-53.
- Deming, W. E. (1943). *Statistical Adjustment of Data*. New York John Wiley & Sons, London.
- Dillman, D. A. (2002). “Presidential Address: Navigating the Rapids of Change: Some Observations on Survey Methodology in the Early Twenty-First Century.” *The Public Opinion Quarterly* 66(3): 473-94.
- Lenhart, A., J. Horrigan, et al. (2003). “The Ever-Shifting Internet Population: A New Look at Internet Access and the Digital Divide.” Washington, D.C., Pew Internet & American Life Project.
- Spooner, T., P. Meredith, et al. (2003). “Internet Use by Region in the United States.” Washington, D.C., Pew Internet & American Life Project.
- U.S. Dept. of Commerce, Bureau of the Census. CURRENT POPULATION SURVEY, AUGUST 2000: INTERNET AND COMPUTER USE SUPPLEMENT [Computerfile]. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 2000. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2001.