How long Can This Go On? Panel Sustainability in Measurement

Tony Incalcatera, Jo-Ann Romano, Maria Cole – Ipsos Connect

Introduction –

In 2015, after more than a year of intensive testing, the Ipsos Affluent Survey USA shifted its data collection from a mail-based survey to online using panel-supplied sample. In that time, we have become more attuned to the unique challenges facing researchers seeking to conduct syndicated research using panel samples.

One of the most important benchmarks of syndicated research is the need to be able to replicate the methodology over time. In this way, syndicators can feel confident that results are reflective of market changes and not the effects of methodological variation. Given the nature of panel sample, it therefore requires that we use qualified partners who maintain high standards for recruitment and member maintenance. In addition, it is incumbent on the syndicator to monitor the quality of the respondent data that is collected to ensure that there is strict adherence to the established specifications.

This paper will describe the development of the procedures and the ongoing quality checks necessary to ensure a continuing and successful syndicated study. In addition, it will present some topline results that will show that it syndicated research conducted in this fashion can produce reliable and trendable results.

The ecosystem of panels –

Panel sample providers abound. Ipsos, which owns and operates its own panel, the Ipsos ISay panel, works with more than 250 other panel companies and regularly evaluates new potential partners. Established companies, such as Research Now and SSI, mingle with newer companies, such as ThinkNow and Epitome, in what might seem like an unlimited supply of fresh sample.

Despite targeting one of the hardest to reach audiences – upper income individuals – through 3 years of almost constant interviewing, we have yet to run into a lack of sample. However, mergers and acquisitions create a constantly evolving landscape revealing that sample can more limited than we would like.

Coupled often with somewhat opaque business practices masking the "true" size of panels, retention rates, and recruitment efforts can make reliance on panel sample a scary proposition for audience measurement if a strong quality control system of monitoring isn't put into place. It goes without saying, that before undertaking a sustained program of interviewing, syndicated researchers should conduct thorough evaluations and testing of their potential panel partners.

For example, a panel company may tout in their marketing materials that they have 13 million members globally. However, upon further examination the number of active members (e.g., those who have responded to a survey in the past 6 months) may only be a fraction of that. The others may have lapsed due to any number of reasons (e.g., waning interest in taking surveys, few interesting surveys being offered to them, life stage changes).

Many panel companies offer a wide range of potential sample, covering virtually every facet of the population. At the same time, there are numerous panels that specialize in delivering more narrowly-defined populations, such as young men, Hispanics, African-Americans, etc. It is important to understand the recruitment policies that panel partners use, the general make-up of the sample attained from these vendors, and ultimately whether the panel will potentially bias results due to skews within the sample. As an industry, we've known for some time that the use of multiple panels, rather than a single-panel ameliorates the concern about single panel skews.

To complicate matters, sample companies are further segmented into those that have a direct relationship with the panelist and companies that act as aggregators of smaller panels, offering these smaller companies the ability to scale up to meet the sample requirements of clients. The actual process in which the panelists are invited to participate in a survey also vary, with some companies utilizing an active direct email invitation to the respondents, while others passively post available surveys and allow panelists to select the survey(s) they would like to complete. Still others rely on a router or river system – a hierarchical screening system to direct panelist to surveys that are most appropriate for them.

The Ipsos Affluent Survey USA experience –

The evaluation and decision to move forward with a reliance on partnerships with panel companies to field the Ipsos Affluent Survey USA (IAS) hinged on several factors. We needed to be sure that the overall industry could continuously provide a pool of respondents that, over time, was:

1) Large enough
2) Diverse enough
3) Refreshed on a reasonable schedule

In addition, we had to feel assured that we could find partners who would properly screen and monitor their panelists and adhere to our strict requirements of continuously fielding surveys at predetermined levels.

Throughout our discussions with potential partners, we stressed the long-term nature of the project, since the program would need to go on in perpetuity. A great number of companies were dropped from consideration since they were likely to run out of fresh sample within a "short" period of time. Although most of the potential partners were engaged in continuous tracking programs, few were experienced with the triple-threat that we presented. Our survey was/is:

1) Narrowly targeted – meaning that only adults living in households with high incomes were eligible
2) Long and conducted in two parts (so it involved a re-contact of respondents 3 to 4 weeks after completing the initial survey)
3) Required that each partner contribute a set number of completed interviews within an 18-quota group structure on a weekly basis

To be sure, we recognized that there would be some give-and-take; perfect symmetry for this plan was impractical, however the variation had to be within reasonable limits to guarantee that the data collected was not subject to bias introduced because of season/time. In other words, we did not want one individual time-period's results to skew overall information because of an over-abundance or dearth of interviews conducted.

Identification of partners –

As we set out on our journey, we first needed to identify and recruit a significantly large group of third-party sample vendors. Each of these vendors needed to pass our rigorous testing and prove capable of adhering to our strict methodology. We have a team at Ipsos who are tasked with recruiting third-party panel sample vendors for all the online interviews the company conducts.

Each of the vendors that were approached had already undergone a rigorous vetting that assessed the supplier's sample capabilities and quality procedures. This included questions about the vendor's own internal sampling methods and data quality measures, their survey invite frequency, their rules for eliminations, and the types of incentives they use.

After determining that the vendor's operations are in line with Ipsos' standards, they are then registered in the Ipsos Cortex database. Cortex is our survey routing system that allows vendors to send sample into our surveys and monitors returns 24 hours a day/7 days a week. The vendor is then tested, and scored, on their actual capabilities.

Once vendors are approved and brought onboard, this team monitors the daily sample returns and, among other things, keeps the vendors on track if their returns are low or inconsistent with their feasibility estimates.

Before a project begins, approved vendors are briefed on the specifics of the survey the panel will be used for. Topics include CPI costs, quota targets, LOI, # of needed completes, field time, incentives and re-invites.

Prior to the start of testing in 2014, the internal Ipsos team sent out a request for information (RFI) to identify which of the numerous panel companies Ipsos works with could supply sample to the study according to the specifications needed. We provided exacting details including, but not limited to:

1) Expected incidence
2) Length of survey
3) Types of survey questions
4) Respondent exclusions required (e.g., no media company affiliations, length of time before a respondent could retake the survey)
5) Requirements to identify and remove cheaters, speeders, and other survey miscreants
6) Anticipated costs

The IAS sample plan involves specific quota targets defined by a combination of gender, HHI and age breaks that require feasibility estimates from all participating vendors prior to going into field. We are targeting these percentages to be as close to the Current Population Survey (CPS) of the Bureau of the Census estimates for the same groupings and want to make sure we have enough coverage for these targets among all of the vendors combined.

Over the course of a year, the IAS is conducted in two six-month waves of data collection. There are semi-annual releases that report on 12-months of data, employing a rolling average of results, whereby the oldest six months of data is replaced by the newest six months. Third-party sample vendors are assessed yearly and as new sample vendors become available, they are set up for testing. It is our mission to cast as wide a net as possible to assure we meet the IAS minimum number of completed surveys requirement (12,500+) each wave.

The current IAS wave uses 10 different panel survey providers. In addition, we are testing offline five previously unused vendors for possible inclusion in the next wave. The testing will allow us to determine if any new vendors can be integrated into the survey protocol, and to what extent.

Training the partners (internal and external) –

Given the economics of panel management, most panel companies prefer to work on faster turnaround projects. They want to launch invitations, gather completed interviews, close out the survey and move on to the next project. In this way, they can utilize their panels and their staff time most efficiently. The problem is that this type of protocol bunches up interviews into a

short period rather than spreading them evenly across time. Knowing this, the IAS team spent a great deal of time setting out procedures for how the interviews should be conducted and then training both our internal project management team, as well as the vendors' teams.

Training is something that unfortunately must be regularly repeated, as personnel changes and complacency settle in. When new people (internal and external) take on new responsibilities, they are usually unfamiliar with the needs of a syndicated research study and therefore must be taught why the protocol needs to be followed as closely as possible to ensure quality.

Panel companies that miss targets, fail to follow schedules, or are otherwise noncompliant face elimination as a sample supplier. Since we want consistency in our fielding, we work hard to stay on top of the companies missing their targets to identify and correct as quickly as possible any issues that come up. These issues may be technical in nature (e.g., survey links that are broken), caused by miscommunication (e.g., project managers relaying the wrong instructions), or something of unknown cause. Our staff has been trained to look for anything out of the ordinary in order to circumvent problems before they affect the survey.

Design meets reality –

The ideal goal of our sampling strategy is to conduct interviews in same proportion that our 18-quota groups exist in the most current U.S. census data from the Current Population Survey. This would enable our sample balancing and weighting applications to be contained within tight ranges and allow us to accurately project results to U.S. Census statistics. Due to the difficulties of reaching certain segments of the population (particularly young men), an exact match proportionately was impractical, so allowances were made that allowed us to remain with reasonable sample balancing factors.

Ultimately, we chose a group of sample partners that mixed larger providers with smaller independents in order to produce a blended sample that would achieve our targets. The decision was based on overall feasibility. Only three of the 10 used in this wave can regularly produce completed interviews across each of the 18-quota groups. The remaining seven commit to completed interviews in 11 to 17 of the groups and Best Effort (BE) for the remaining. The best effort contributions are welcome since they help close the gap between feasibility and the proportionate goal.

The table below (Table 1) is an example of how our quota structure is designed in advance of beginning a new wave of fielding.

Table 1 – Example of Sample Feasibility

Wave Quota-Group Feasibility Example	Goal	Feasibility (Weekly completes)		Expected contribution per week									
	Proportion based on Current Census Data	#	%	Panel A	Panel B	Panel C	Panel D	Panel E	Panel F	Panel G	Panel H	Panel I	Panel J
Male $125K-$149K Age 18-29	3.4%	11	2.1%	1	BE	1	BE	1	1	2	3	2	BE
Male $125K-$149K Age 30-49	5.9%	45	8.7%	2	1	2	1	1	3	20	3	11	1
Male $125K-$149K Age 50+	6.2%	50	9.6%	2	2	2	2	1	2	22	3	13	1
Male $150K-$249K Age 18-29	5.8%	11	2.1%	1	BE	1	BE	1	2	2	3	1	BE
Male $150K-$249K Age 30-49	9.6%	40	7.7%	2	1	2	1	1	3	10	13	6	1
Male $150K-$249K Age 50+	11.1%	49	9.4%	2	2	3	2	1	2	13	13	10	1
Male $250K+ Age 18-29	1.8%	6	1.2%	BE	BE	BE	BE	BE	2	BE	2	2	BE
Male $250K+ Age 30-49	3.7%	17	3.3%	1	1	1	1	BE	3	2	3	5	BE
Male $250K+ Age 50+	4.3%	20	3.9%	1	1	1	BE	BE	2	5	2	8	BE
Female $125K-$149K Age 18-29	3.0%	14	2.7%	1	BE	1	BE	1	1	3	3	4	BE
Female $125K-$149K Age 30-49	5.6%	57	11.0%	2	1	2	3	1	2	27	3	15	1
Female $125K-$149K Age 50+	5.9%	52	10.0%	2	1	2	4	1	2	23	3	13	1
Female $150K-$249K Age 18-29	4.9%	13	2.5%	1	BE	1	BE	1	2	2	4	2	BE
Female $150K-$249K Age 30-49	9.4%	46	8.9%	3	2	3	2	1	2	14	11	7	1
Female $150K-$249K Age 50+	10.4%	48	9.2%	3	3	3	4	1	1	13	11	8	1
Female $250K+ Age 18-29	1.6%	6	1.2%	0	0	0	0	0	2	0	2	2	0
Female $250K+ Age 30-49	3.6%	17	3.3%	1	1	2	1	0	2	3	3	4	0
Female $250K+ Age 50+	3.8%	17	3.3%	1	1	1	1	0	1	4	3	5	0
Total	**100.0%**	**519**	**100.0%**	**26**	**17**	**28**	**22**	**12**	**35**	**165**	**88**	**118**	**8**

The role of 24/7 Monitoring –

Once the IAS is in field, extensive quality procedures and standards are in place throughout the field period that are reviewed diligently by the Ipsos IIS department to ensure respondents are "real", "unique", "fresh" and "engaged".

- "Real" respondents: Checks that respondents are who they say they are using various methods of detection for robots, data anomalies, device setting and geo location.
- "Unique" respondents: Checks that respondents can take the survey only once by identifying duplicate emails, devices and contact details.

- "Fresh" respondents: Monitors that respondents have not participated recently in similar surveys via strict panel usage rules in addition to what is already stated above.
- "Engaged" respondents: Checks for straight-lining, speeding, and the removal of "bad" or inactive respondents.

The Ipsos project management team performs their own series of checks and balances to monitor the panel sample quality before, during and after the survey is in field:

- Target totals: Weekly targets are set prior to fielding. Some targets are based on vendor feasibility estimates while others are based on the current U.S. census estimates. The number of surveys marked as "complete" are checked to ensure they are in line with the targets that were set. Any significant deviation in the number of completes when compared to the original vendor estimate would result in immediate discussions with the vendor to look for any technical issues with the sample outgo process or other possible reason for the discrepancy, while all IIS internal sample settings and procedures would be re-evaluated.
- Other return codes: The total number of abandoned, fraudulent, screen-out and quota full records are assessed and any totals that fall outside of what is considered reasonable are investigated until a cause is discovered or could be logically explained.
- Length of Interview (LOI): The median length of time to complete the survey is monitored. Testing is done internally on the survey prior to fielding to check not only that the programming is correct but how long it should take, on average, to complete the survey. If the median LOI while in field deviates significantly from the results of the testing, especially if the LOI is much shorter, then a closer look would be taken at the data itself by the QA team and the records of the respondents who are deemed as not "engaged" as defined in the section above would be dropped.
- Attempts-to-Completes ratio: The ratio of the number of completed records to the number of overall attempts (all returns) is compared daily to the results from all previous days to check for patterns or anomalies.
- Vendor-specific assessments: The total number of completes on a vendor by vendor basis is charted every few weeks and lined up against the results from all prior waves for that vendor based on the same point in time (e.g., 4th or 16th week in field). If there are differences that cannot be explained by factors such as reductions or changes in feasibility estimates, discussions are undertaken to find the reason for the drop-off (or surge) in results.
- Data cleaning: On-line surveys are set up to cut down or eliminate completely the amount of data processing necessary. The Ipsos Affluent Survey employs an additional step in its monitoring of the quality of the panel sample by running all completed survey data through a detailed data cleaning program to ensure that all question responses are valid and that all skip pattern and other special instruction logic is programmed correctly.

Although it is important and necessary to employ standards and procedures to evaluate and recruit reliable, qualified third-party vendors and then monitor and assess the panel sample quality with the ultimate goal of sustaining on-going panel-based syndicated research, it is also an established practice by the Ipsos sample team to develop a solid working relationship with the contracted third-party sample vendor companies. This helps to instill a sense of trust on both sides and keeps open the lines of communication and cooperation, thus avoiding major pitfalls than can occur when there is tension between parties or hesitation to contact the other side if there were critical questions that needed answering.

In the two-and-a-half years that we have been conducting the survey online, the vast majority (84.8%) of all interviews have come from sample vendors that have been with us from the beginning. Four of the original vendors were eventually eliminated after their returns diminished over time, and two vendors passed initial tests, but failed to produce reasonably and they and their respondents were purged from the study. As mentioned earlier, one vendor was acquired by another and their panelists integrated into the acquirer. To replace eliminated vendors, we tested and brought on board three panel sources that continue with us today. The pretesting was done to ensure that changes would not have a deleterious effect on survey results.

Results –

The survey has been running continuously since January 2015, so we now have two and a half years of results. Study releases are semi-annual (Spring and Fall), using a 12-month rolling-average, whereby the oldest six months of data is replaced by six months of new data. Comparisons are made Spring to Spring and Fall to Fall, rather than Spring to Fall or Fall to Spring. This eliminates the smoothing result of six months of common data.

Chart 1, below, shows the correlation between our Spring 2017 and Spring 2016 releases using our Total Brand Footprint (TBF) measurement. TBF is the sum of all exposure that adults covered by the survey have with the 290+ brands that are reported. The Spring to Spring comparison produces a correlation coefficient of 0.9948.

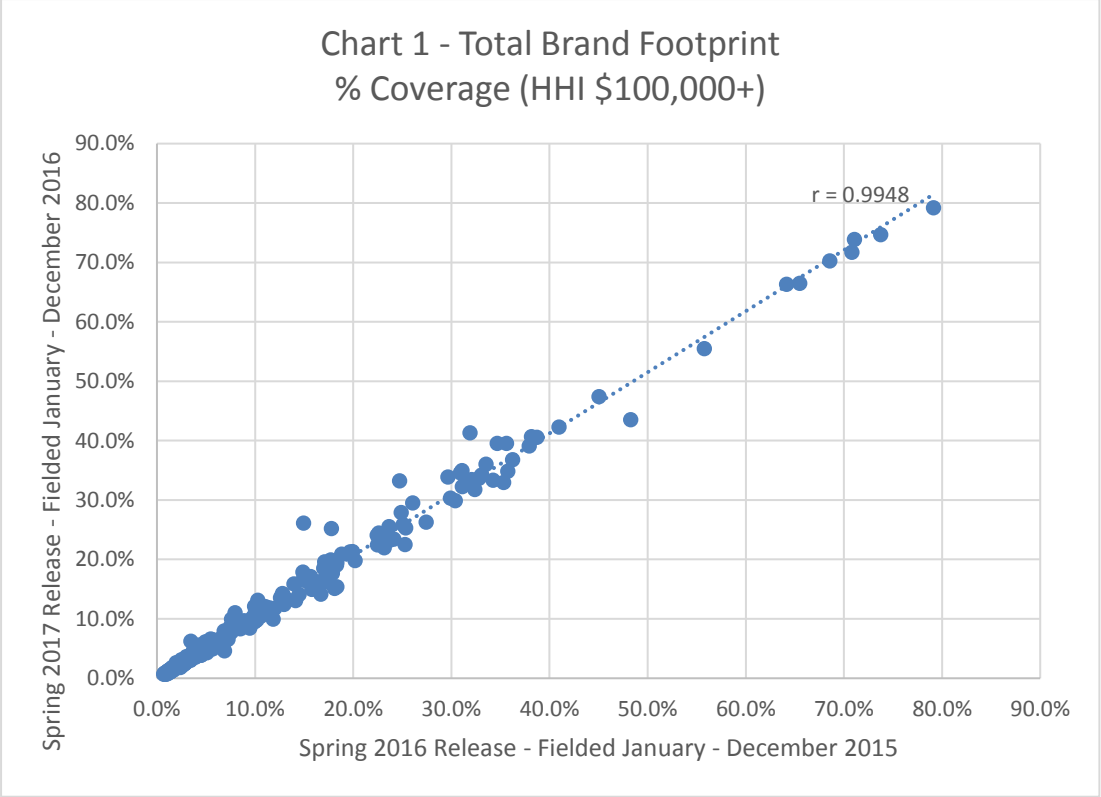
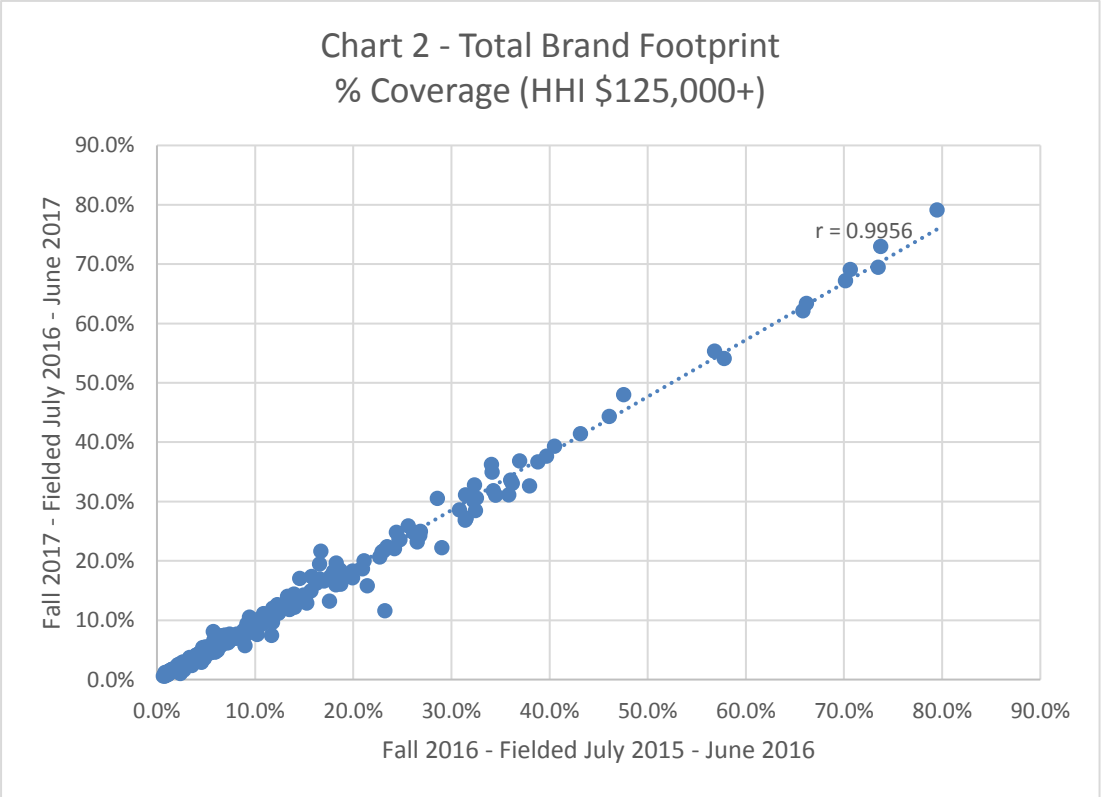


Chart 2 shows the same type of analysis, comparing our Fall 2017 data to our Fall 2016 data. It should be noted that as household incomes have continued to grow in the U.S. we found it necessary to increase the minimum income requirement for inclusion in the study. Effective with the Fall 2017 release, the minimum threshold for inclusion was $125,000 or more in household income as opposed to the previous requirement of $100,000+. Consequently, the comparison in Chart 2 is based on adults with HHI $125,000+ in both study releases.



The correlation is slightly higher in the Fall to Fall comparison (0.9956). The largest outlier is a media brand whose name changed dramatically over the two-year period while refocusing its target market and programming.

Variance Estimates and the Design Effect as a Measure of Stability –

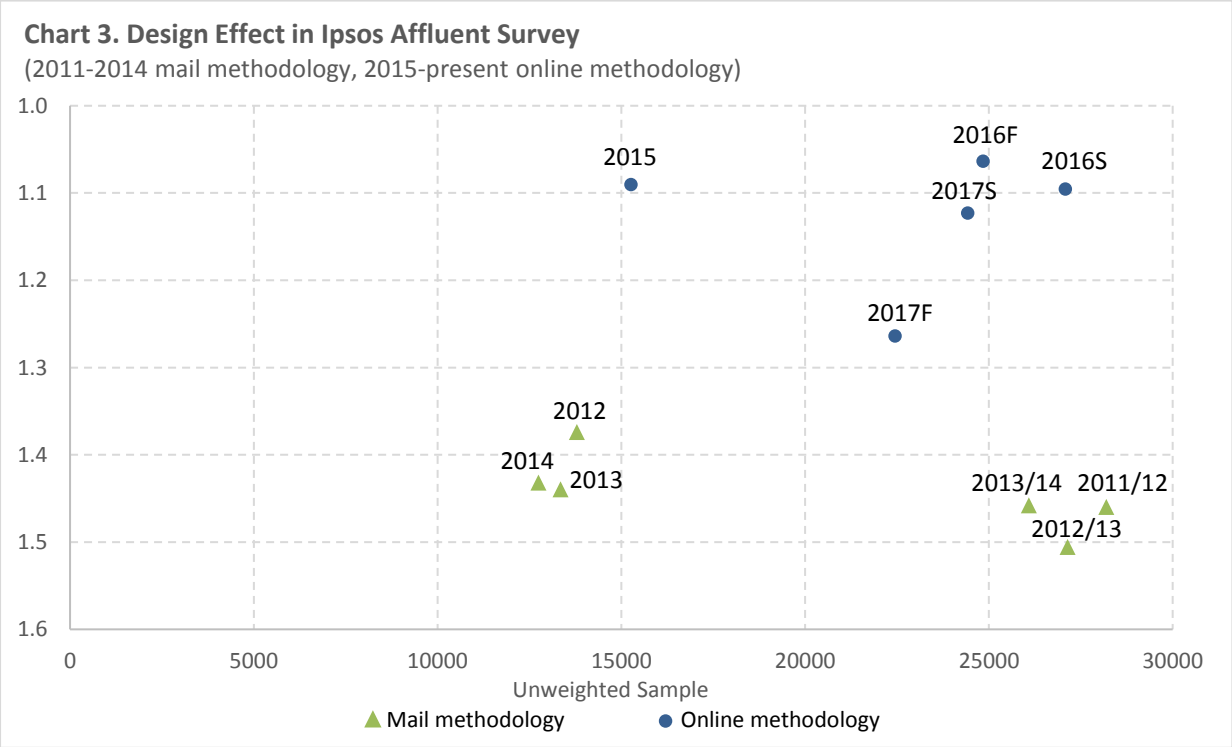
The sampling methodology discussed previously means that we have a survey with complex design, as opposed to one coming from a true simple random sample (SRS). While we have been able to greatly improve the alignment between our sample's demographic characteristics as compared to population estimates (per Census data) since moving to online data collection, variations remain that necessitate post-stratification weighting. When estimating sample variances, we cannot use the straightforward approach one would use under SRS conditions, but instead must make upward adjustments to the variance estimator to account for the complexity of the design. The design effect (DEFF) is the ratio of the variance under a complex survey design to the variance under SRS conditions. In our study, we have been using a jackknife estimation to determine the DEFF, which is then used in adjusting the two-sigma tolerance levels clients use to estimate confidence intervals.

When comparing historical values of DEFF in IAS, we note that they were much larger under the old mail methodology than now with our online survey. A DEFF value of 1.0 means there is no design effect--variance estimates are the same as would be expected under SRS conditions. Between 2011 and 2014 (mail studies), DEFF varied from a low of 1.37 to a high of 1.51. Since 2015 (online studies), DEFF vary from a low of 1.06 to a high of 1.26. The higher DEFF associated with the Fall 2017 release is largely caused by the transition to the new income threshold mentioned earlier. During the first part of the transition we continued to interview adults with household incomes between $100,000 and $124,999 to keep the studies trendable from one to the next.

While sample size plays a role in the relative DEFF, we note it is not a contributing factor here. Under the mail methodology Doublebase samples had higher DEFF with higher unweighted sample sizes than online methodology releases with smaller sample sizes (see chart 3).

The smaller DEFF observed under online data collection is indicative of a more stable study design and less-biased variances. The main driver of the decreases in variance estimates is an improvement in the statistical efficiency of the sample. What this means is that the range of respondents' weights is smaller, meaning the demographic characteristics of the unweighted sample are more closely aligned to population estimates under the current online methodology, which uses strict quotas as described previously.

In addition to stability in results that we demonstrated comparing total brand footprint data, the overall decreases in DEFF mean that the variances (confidence intervals) around media and all other estimates are smaller, providing for more reliable data year to year.



Conclusion –

Based on our experience, we believe that panel sample will continue to be a viable option for syndicated research. In our case, we are only interested in interviewing the top fifth of all adults in the U.S. based on household income. Consequently, syndicated researchers looking to conduct surveys amongst a broader population should have no difficulties finding sample.

Finding adequate sample, though, is only part of the equation. In order to ensure data quality, rigorous systems to continuously monitor progress and all return totals are required. Complacency could result in procedural errors that will only multiply over time if left unchecked, with the potential of undermining efforts for a successful research study. Foresight and vigilance are required elements of any syndicated study, but they become even more critical when using panel sample. There are too many variables that can cause aberrations in the data if left unchecked.