Using Measurement Science to Link Traditional and Emerging Research Methods with Calibration Panels

Pasquale (Pat) A. Pellegrini, Ph.D., Steven Millman, Tamara Barber & Bill Harvey*

Simmons Research, NY, NY.
***Strategic Advisor to Simmons Research**

Introduction

Market researchers are increasingly enjoying access to a multitude of natural occurring data or data "in the wild" meaning data that are from transactions-based sources, like merchant point of sale purchase data. Typically, natural data provide massive scale, low latency or even immediacy of data reporting, and a strongly standardized data collection mechanism. Transactions-based information are data culled as an artifact of a transaction like a credit card transaction is an artifact of a system designed specifically to create a communication between a card holder, their bank, the merchant, the merchant's bank, and nothing about that transaction has an element of statistical or consumer research in it. Nevertheless, transactional data, when properly adjusted, edited and weighted, can provide valuable and actionable insights about the consumer, and potentially decrease the burden of active consumer data collection by survey research firms such as Simmons (Palet and Engel, 2016).

Market researchers also have access to passively collected data that uses measurement technology to capture various behaviors as consumers go about their lives, be it shopping, online and mobile browsing, or watching TV and video. As with transactions-based data, these collection methodologies can provide massive scale, low latency data reporting coupled with highly standardized data collection mechanisms. However, these data are inherently biased in that they are not balanced back to a representative and projectable population since researchers leverage the measurement technology to maximize measurement across the variables with the highest variance requiring massive convenience samples. For example, in the media world, passive measurement of TV/video consumption is designed to capture the hundreds of programs, stations and devices capable of delivering linear, non-linear and over-the-top (OTT) programming. Researchers interested in TV/video use both transactional data (cable/satellite STB data) or passive collection from Automatic Content Recognition (ACR) applications on smart phones/Smart TVs/connected devices or passive meters on laptops and tablets.

At Simmons, as we continue to tackle problems around traditional survey sampling methodology like sample frames, data collection modes, and response rate issues, we are also heavily engaged in the development of methodology to deal with newly emerging transaction-based data sources. While we have plenty of experience with the integration or fusion of advertiser first-party data with high quality survey data (see description of NCS below), the scale and complexity of emerging transactions-based data demand the development of new techniques. The mining of transactions-based data for the purposes of collecting insights on consumer behavior requires rigorous research and validation. The term we use is Measurement Science: the application of measurement methodologies (sample design, weighting, calibration and validation) and measurement technology (including the collection of passive, massive-scale, granular data or natural data) to develop datasets that link often imperfect, fast-moving data to a higher quality, projectable sample data.

For TV/video consumer behavior, some practitioners have chosen to continue to use traditional probability samples such as in the case of Nielsen, while others have used naturally occurring data in conjunction with non-probability samples e.g. comScore/Rentrak (now known as comScore TV). The research reported here starts from the premise that all data can be leveraged for its unique value, by grounding all types of data using a calibration sample that is the "truth standard" to which the other data types can be calibrated. This calibration sample must be a high-quality probability sample and over time its size can be increased. Questionnaires over time can be shortened by replacing data with naturally occurring passively collected data. In this paper, we describe some preliminary work Simmons has done to set up the groundwork for carrying out this vision, namely using non-probability TV/video data with high-quality consumer and attitudinal data, and specifically trying to identify which variables need to be measured actively or passively and used in calibration.

Background

Using some form of calibration panel to develop calibration weighting variables to balance non-probability data is not new to the market research world, albeit not widely used until only recently. Over 15 years ago in the US, comScore pioneered the use of an offline, random-recruited calibration panel to develop behavioral universe estimates to generate weighting targets for "time spent online" that were used to "control" the larger, convenience sample recruited Media Metrix panel. Pellegrini and Meierhoefer (2011) discussed the role of both probabilistic reference points for massive scale non-probabilistic panels and validation using data intensive techniques like Jackknife Replication.

Calibration weighing was tested by DiSogra and his colleagues at GfK (DiSogra, 2011) where data from different sources was combined by using estimates from one source as "benchmarks" to calibrate the data. This research was highly pragmatic as the large internet based Knowledge Panel (n=55,000) still has limitations in sample size for small or rare populations (e.g., rare medical conditions, specific race/ethnic groups, ownership of luxury products). Such approaches may also work where finite sample sizes are an issue based on small geographic area samples, assuming spatial variations can be captured. In these applications, supplemental opt in cases with quota sampling was used to resolve demo skews and extreme weights.

Still, recent research concludes that some of the key challenges to doing this well, include clearly outlining and testing any assumptions about the specific non-probability sample. In this vein, RTI tested an approach to balancing survey data that had two different sample frames: Facebook and an address-based sample. They tested an assumption that the demographic characteristics of address-based respondents who took the survey online would be the same as those who were recruited via Facebook (Kott, 2017).

Finding ways to include calibration variables beyond demographics in each dataset to be integrated, Australia's Social Research Centre undertook research that explored the pros and cons of four calibration techniques that were used to combine surveys across a probability sample with a non-probability online panel. The research showed that – while there are no one-size-fits-all solutions – the scope of variables available for calibration must include enough sample and enough variables in common across datasets (Nieger, et.al., 2017).

In a collaboration with Simmons Research, RMT's DriverTags™ technology were tested as possible calibration variables in explaining usage of thousands of brands measured in Simmons Research surveys. DriverTags™ are psychological dimensions of television/film/video content including ads as well as programs, reflecting the human values expressed, the character strengths and weaknesses of the people in the content, moods/emotions evoked, and all other attributes that could account for the motivations of the person who is attracted to viewing that content. It was found that these variables add explanatory power even on top of extensive demographics, geographic variables, and Simmons' current approximately 700 attitudinal questions.

Bill Harvey and his colleagues at TRA and Next Century Media (NCM) have extensive experience with calibration research. TRA (now called TiVo Research since TiVo acquired TRA in 2012) showed that Nielsen ratings could be closely approximated by projecting Set Top Box data in an iterative rim weighting system conforming weights to thirteen demos and 80 geographic strata. Dating back to 1997, NCM collected the first research grade STB data and published a false positives algorithm that corrected for the tendency of people to leave their STBs powered on when they turned their TVs off, and this algorithm resulted in Homes Using Television (HUT) data closely conforming to Nielsen data by half hours of the day (Harvey, 2004). Using STB based HUT to predict PUT levels was the focus of work by Pellegrini (2006). The calibration of transactions-based data like STB or other video return path data across platforms is of considerable interest as measurement of consumer video consumption continues to advance, but is complicated by increasing fragmentation across video enabled devices. The complexity of TV/video measurement and the need for calibration is the subject of this paper.

TV/Video Calibration

In an ARF conference paper, Palit and Engel (2016) reported on the validation of a calibration technique to correct for potential biases in viewing levels derived from STB data for comScore TV. The authors note that there is a need for data adjustments in all transactions-based data sets because of the presence of unknown sources of systematic bias that may be observed, but not identified. For example, they report on a case where a data "leak" occurred meaning that an observed drop in TV viewing hours was the result of some unknown change in the data collection system by one or more of the third-party providers (or MVPDs, Multichannel Video Providing Systems) to Rentrak. A correction factor is applied using historical viewing levels per household from the prior year and is validated using published person viewing levels from Nielsen.

comScore TV is but one example of "hybrid" TV/video measurement. The ASI TV Symposium for several years now has been reporting on the progress on combining viewing data gathered by research panels with server, STB or transactional data by leading audience measurement providers around the world such as Kantar, GfK, and Nielsen. In the U.S., solutions to the increasingly complex world of TV/video consumption has led media giants like Viacom, Turner and NBCU to bring together data from various sources. The goal, of course, is to follow the consumer consumption from linear TV, to DVR playback, Video On Demand (VOD), and digital streaming or over the top (OTT) and even to mobile. Hence, measurement solutions that embrace multiple data sources, from panel based samples to transactional and passive meter based convenience samples, are essential.

The research presented in this paper explores a specific challenge of combining TV/video research sample data with transactional passive data from a convenience panel. The Simmons Research National Consumer Study (NCS) serves as the benchmark of US adults' demographics, psychographics and key behaviors, which allows us to inform a data integration with passively collected linear TV, DVR, VOD, and OTT viewing behavior. The opportunity to do this work arose from the Symphony Technology Group (STG) ownership of both Simmons and Symphony Advanced Media (SAM). SAM has subsequently been discontinued, but while it was running it collected information using a downloadable mobile app that elegantly measured all types of television/video exposure passively and from a single capture system. The system was executed using a non-probability sample and our hypothesis was that SAM would produce audience viewing estimates with deep consumer insights if SAM could be calibrated to Simmons' probability sample.

The NCS consists of a representative, projectable probability sample of 25,000 U.S. respondents annually, covering over 60,000 consumer attributes including cross media consumption. For this research, we required extensive demos, a full suite of psychographics, and self-reported TV viewing at the network and show level within the past seven days and past four weeks. The SAM data came from their Video Pulse OTT measurement product that used a convenience panel of 15,000 panelists who opted-in to having their devices "detect" video content they view. Through a proprietary technology, SAM tagged not only what shows panelists were watching, but also inferred the device (TV, Computer, Tablet, Cellphone) and mode (Live, DVR, VOD, OTT) of the viewing behavior. Data were collated at L+3, L+7, and L+35 timeframes among others, so that it aligned with third party currency metrics already used in the market. (L+3 means viewing the program live or time-shifted within 3 days after the live telecast, and so on for the other timeframes.)

Research Objectives and Methodology

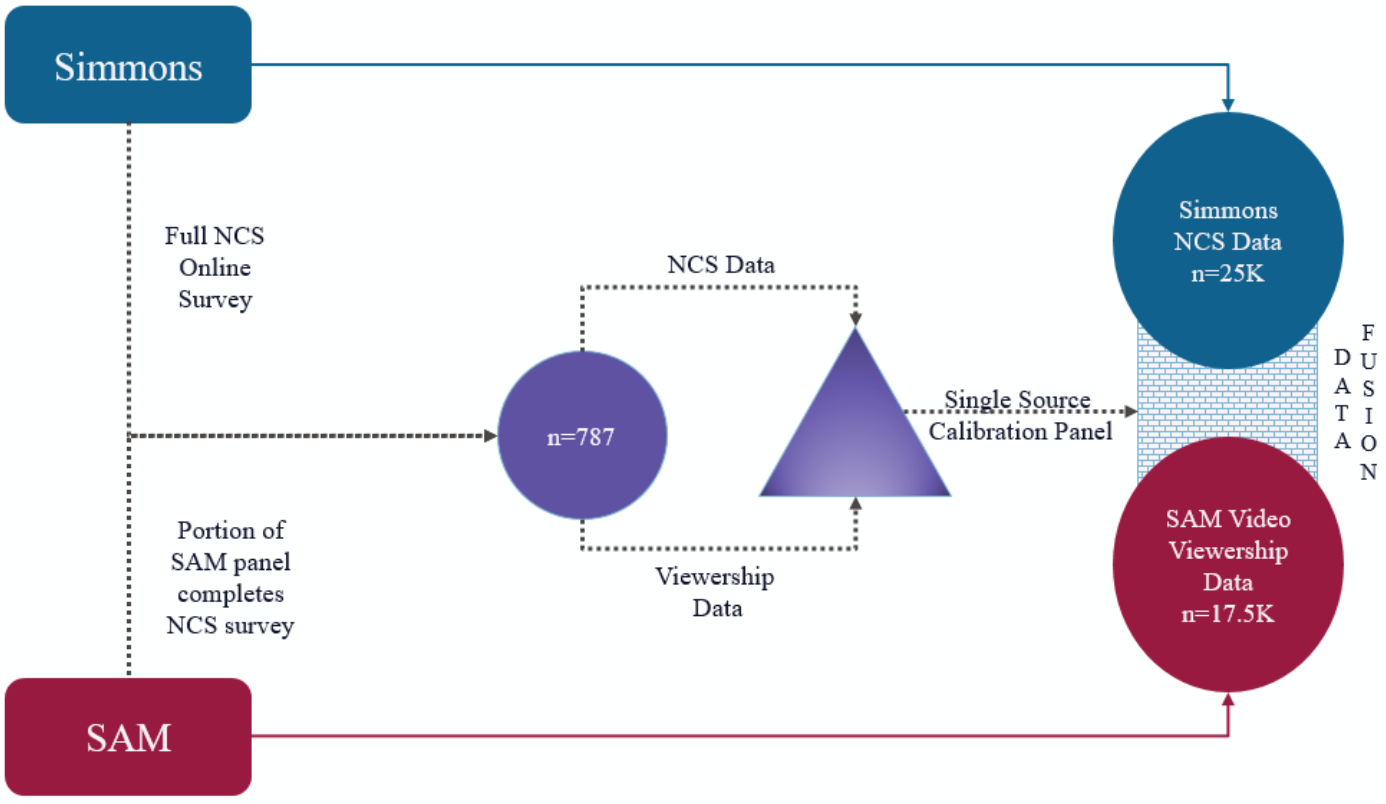
An estimated 28% of broadcast and 26% of cable viewing happens outside the universe currently reported by currency measurement – i.e. TV ratings. Measuring linear, traditional TV viewing is no longer enough for advertisers and networks to understand the true reach of their content. Although various research vendors are making inroads into measuring some non-linear TV viewing by capturing DVR out to 7 days and VOD out to 3 days, etc., there does not yet exist a currency data source that captures the full audience that watches TV programming across Linear, DVR, VOD and OTT out to an extended period e.g. 35 days. Advertisers lack a clear line of sight into how well OTT media buys will reach their target audiences. The goal of this research was to advance the understanding of this currently under-reported behavior, by developing analytics for OTT viewing that are projectable and balanced to the US population and linked to the myriad attitudinal, preference and brand data available in the NCS.

While much of the research around calibration discusses bringing non-probability data into line with the "truth" based on probability samples, our approach starts with a statistical fusion of the non-probability SAM panel data into the probability sample of NCS respondents. This approach allows us to validate OTT and linear viewing levels and ratings overall, and more importantly it allows us to take full advantage of the extensive NCS psychographic, brand and behavior measures across a multitude of categories. Paramount in this research, and consistent with our measurement science philosophy, is to develop the capability to layer in insights about what drives viewing behavior, or characterizes different types of viewers across viewing modes by combining Simmons and SAM data. Internally to Simmons, this project became known as "video analytics" and remains the methodology behind a future product tentatively named Video Viewer Insights (VVI).

Before setting out on the fusion itself, we explored to what extent the SAM data tracked with publicly available currency (Nielsen) ratings across 299 shows. To this end, we found that there already existed a .9 correlation between SAM's Live rating and Nielsen's. An immediate goal, then, was to improve these correlations, the assumption being that the closer the fusion data is to Nielsen ratings for Live TV, the more validity this would give to the OTT rating coming out of the fusion and thus provide a strong basis for VVI.

To gain the most knowledge out the fusion exercise, we first needed to develop an overlap sample to serve as an initial "calibration panel" where the same panelists were directly connected to both SAM viewership attributes and NCS self-report attributes. This single-source calibration panel was randomly selected from the SAM convenience panel from December 2016 and January 2017 and provided 787 respondents that then completed an online version of the NCS. In an ideal situation, a single-source calibration panel would be recruited from the high-quality probability NCS panel, and Simmons is currently working on such a panel as part of an extended Measurement Science based product portfolio. For this paper, the so-called overlap calibration panel became the basis for learning for the fusion research work, allowing us to identify: 1) differences between the two standalone datasets, 2) which NCS variables perform best in predicting SAM viewership behavior, and 3) the drivers of differences between how Simmons and SAM correlated with Nielsen Live ratings. The calibration panel allowed us to see the actual overlaps between SAM and NCS metrics as the most important basis for evaluating the goodness of various types of fusions of SAM data onto NCS. Eventually, such an exercise will help us understand where and when a calibration panel would be required, what it needs to specifically measure, and any other key specifications as building such a panel could be prohibitively expensive. Figure 1 shows the calibration process at a high level.

Figure 1. The Simmons and Symphony Advanced Media calibration process (aka "video analytics" or VVI).



The fusion itself was executed using a donor-recipient model executed with the Statmatch software system. We then chose to develop three different models of the fusion based on three different classes of linking variables as outlined in Table 1 below:

Table 1: Fusion model variations			
	Fusion 1 "Demos"	Fusion 2 "Shows"	Fusion 3 "DriverTags"
Key Drivers	Gender + Age + Region	Gender + Age + Region + TV Shows TV Shows based off of factors derived from 200 Primetime network shows	Gender + Age + Region + DriverTags DriverTags based off of 50 clusters derived from 265 DriverTags

Results

As a first test of the models, we assessed the SAM donor rate in the fusion, considering this a measure of how efficiently the donors were being used in linking to NCS respondents. Across the three models, the Demos fusion performed best on this measure. This model used 71% of SAM panelists, compared to a rate of 46% and 44% for the Show and DriverTag™ fusions, respectively. Conversely, the internal consistency and accuracy of the models (meaning the degree to which the shows watched by the SAM panelist and the shows watched by the NCS respondent matched) was higher with the Show and DriverTag™ fusions. Intuitively, this was not surprising given that the Show and DriverTag fusions were essentially using shows – or a derived metric based on shows – as the linkage in the fusion. Said another way, the Show and DriverTag fusions had richer linkage variables – behavioral linkage variables - and therefore we would expect that these variables would match more often when comparing rates of matching shows watched in the standalone SAM and NCS datasets. (See Table 2)

Last, since the calibration panel consisted of panelists who had both SAM and the full NCS data, we also ran each model on that panel and tracked to what extent these panelists "self-donated" in each fusion. For the 787 members of the calibration panel, what would happen if we fused their SAM data with NCS respondents in that pool of 787? How many of those respondents would have been fused to themselves? In other words, how many times would a SAM donor match directly back his or her NCS data if we performed the fusion just as we would with the standalone SAM and NCS data sets?

Surprisingly, the results showed that none of the donors from the calibration panel "self-donated" in the Demo or Show fusions. The DriverTag™ fusion was the only one that showed promise on this front, with 103 SAM panelists donating to their own NCS responses. This suggested that the DriverTag™ fusion was of higher quality in preserving internal relationships between the two data sets, but we suspect this might be discounted by the relative lower rate of efficiency in the fusion where StatMatch only selected 44% of SAM panelists (see Table 2). Another interesting point here is that only 50 clusters of the original 265 individual DriverTags™ were used in the fusion for efficiency, but clearly a follow up exercise would likely gain from the additional variance captured in the full set of DriverTag™ data. Any follow up exercise will be likely to replace StatMatch with another model capable of forcing 100% donation or optimizing percent donation against the other goodness metrics.

Table 2: SAM donor metrics across three fusion models			
	Fusion 1 "Demos"	Fusion 2 "Shows"	Fusion 3 "DriverTags"
<u>Efficiency</u> % SAM Panelists used as donors	71%	46%	44%
<u>Accuracy</u> % agreement btwn shows watched between NCS self report and SAM donor	5%	15%	19%
<u>Self Donation</u> Rate at which donors from the calibration panel donated to themselves in the fusion	0%	0%	10.3%

Another means to judge the quality and validity of our fusion models was to look at their ability to increase, or not, the correlation of our ratings with publicly available published Nielsen ratings. Here, the assumption is that the Nielsen ratings were the "source of truth" with which the fusion models needed to align. (See Table 3). We started with a comparison to average live ratings across Nielsen, SAM and the three fusions and the Demo fusion was the clear winner. It produced an average rating of 1.09 across 299 shows – only three percentage points off from Nielsen's comparable average rating. Likewise, among the 25 highest-rated shows per Nielsen, the Demo fusions produced an average rating of 4.67 (See Table 3).

Table 3: Fusion model comparison to truth.

	Average Live Rating				
Number of shows	Nielsen	SAM	Fusion 1 "Demos"	Fusion 2 "Shows"	Fusion 3 "DriverTags"
299 shows	1.12	0.98	1.09	0.7	0.94
25 highest-rated shows	4.86	4.05	4.67	2.97	4.41

The Demo fusion improved the correlation to Nielsen Live ratings – bringing it up to .94 compared to .9 with the SAM data (see Table 4). This correlation dropped when looking only at the 25 top-rated shows. However, on balance, the findings – particularly around donor efficiency and average ratings -- led us to conclude that the Demo fusion was the stronger fusion model for giving us confidence that any OTT ratings coming out of this model would align most closely to real world behavior. Recall that at present, such OTT ratings do not exist. This result also needs to be viewed in the context of the larger reason for this study, that is, to provide OTT video measurement rich consumer profiles and the "why" behind video consumption across devices and channels.

Table 4: Fusion model correlations with truth and SAM ratings

		Fusion 1 "Demos"	Fusion 2 "Shows"	Fusion 3 "DriverTags"
Across 299 shows	Nielsen Ratings	0.940	0.947	0.944
	SAM Ratings	0.996	0.924	0.947
Across 25 highest-rated shows	Nielsen Ratings	0.738	0.798	0.585
	SAM Ratings	0.982	0.705	0.700

To further evaluate the fusion results, we reviewed the demographic profiles of 20 different networks, and compared those profiles across the models, as well as to estimates from publicly available Nielsen estimates plus the SAM and the NCS estimates. The networks evaluated were a mix of broadcast (2), sports cable (4), kid-oriented cable (4), African-American oriented cable (2), and special interest cable related to cooking/food, home, travel, and original programming (8). We focused the evaluation of the models across Gender, Age, and Race distribution within each of the networks, and we determined parameters that would score each fusion as Pass/Not Pass for any network, or whether we required the fusion need further review. Such an evaluation allowed us to get closer to understanding which behavioral measures could be used in future models to capture internal relationships better than the simple fusions built here (See Table 5).

Table 5: Summary of fusion validation on network level across demographic breakouts

Data sets used	Measures compared across data sets
– SAM – Nielsen, publicly available – Simmons Winter 2017 NCS 12-Month Study – Fusion 1 "Demos" – Fusion 2 "Shows" – Fusion 3 "DriverTags"	– Gender – Age (18-34, 35-54, 55+) – Race (White, Black, Other)
Parameters to determine whether a fusion "passes" review	
– If Fusion estimates are within 5 points of Nielsen – If neither the fusion nor the NCS are within 5 points of Nielsen, but the fusion is moving directionally toward Nielsen	

Using the Demo model first as it was the better fusion for preserving Live ratings alone, we looked at how well this fusion aligned with the demographic breakouts. Not surprisingly, results were mixed. On gender, the model showed unexpectedly low representation of female viewers for seven networks, such as Cooking Channel, Food Network, HGTV, and Nickelodeon. Likewise, sports programming and a handful of other media properties comprised six networks that had low male representation. Overall, the gender breakouts among all 20 of the networks we reviewed tended to closely align to a 50/50 or 40/60 gender split, while both Nielsen and the NCS data suggested that many of these networks should have a stronger skew toward one gender over another. We flagged all thirteen of these channels as "under review" for Gender.

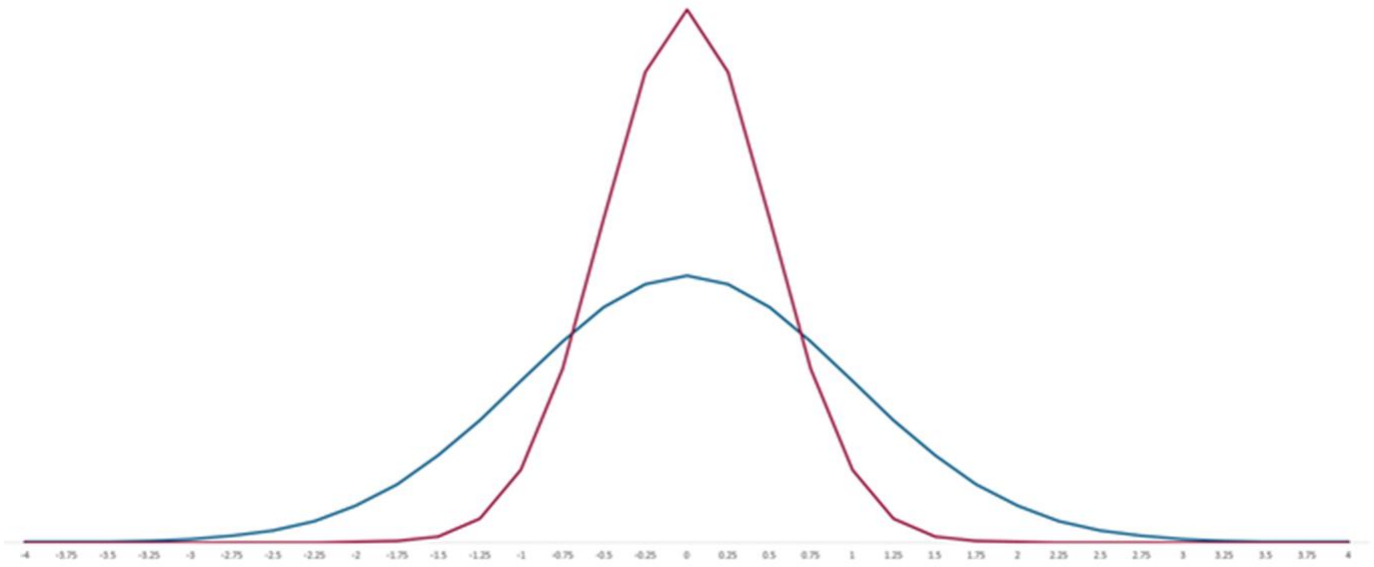
Within age, we found that ten of the twenty networks passed for at least two of the three age groups we were comparing to Nielsen. However, kid-oriented networks and others such as BET, NBA TV and AMC overrepresented viewers age 55+ in the Demo model. Likewise, two broadcast networks had an overabundance of 18-34 year-olds, compared to the age distribution of these networks in Nielsen data. This left ten networks "under review."

The Demo model performed well for race with 14 networks passing for at least 2 of the three race categories. However, African-American viewers were underrepresented not only on the few African-American oriented networks, but also within two sports networks and two kid-oriented networks. Overall, Race within the Demo fusion tended to toward a 70/15/15 distribution for White/African-American/Other, where Nielsen and the NCS data suggested that the audiences should be more diverse in 6 networks that we flagged as "under review" for this analysis.

We repeated the above analysis for the remaining two fusion models and expected that as we started to look past the ratings and look at the demographic make-up of viewers on a network-level, we would see better results given that the two models exhibited greater accuracy in linking SAM donors to NCS recipients who watched the same shows. The results for the Shows and DriverTags fusion models were similar to the Demo model and did not have better alignment with Nielsen or NCS estimates. In the end, the low donation rate provided by StatMatch was probably the key to the disappointing results, and a better multivariate fusion model is needed to capture the richness of the layers of internal data relationships. A second learning to be applied in subsequent methodological research is that the full 265 DriverTags™ should be used rather than a condensation of them.

With a high-quality calibration panel from which to train a fusion, the quality of a match should be much higher. In typical fusions, we start with rule based matches such as men can only be matched to men, southerners can only be matched to southerners, and so on. Beyond these, there are softer matches created based on a series of variables that do not need to be exactly the same between donor and recipient, but rather by the closest overall match usually defined by a "score" such as a propensity score. A calibration panel allows us to take advantage of a training set to create the strongest matches. Without such a training set, we are forced to rely on variables which have a historical or theoretical rationale for being effective, but which may be less effective for the specific data being fused. Figure 2 shows the difference between two such approaches. Since closer scores are a strong match, you ideally would like to see the difference between those scores be as close to zero as possible. In a calibrated fusion (the curve in red), these scores would be much closer because the fusion variables would have been tested and selected for the greatest explanation of variance. In an uncalibrated fusion (the curve in blue), although still unbiased, there would be more variance in scores and therefore more error in the resulting fusion.

Figure 2: Distribution of Propensity Scores for Calibrated versus Uncalibrated Fusions.



Discussion

As the research industry looks for solutions to more efficiently measure consumer and media behavior, for example, TV/video consumption from linear TV to OTT to VOD to mobile, the combining of various survey and transactional/passively measured databases is clearly the way forward. These hybrid approaches involving data fusions have the granularity, low-latency, high volume, completeness (e.g., cross-platform, cross-device, online and offline measurement), but need to overcome the disparate methodologies, technologies, data quality and scales of measurement that persist. If the future of consumer and media measurement is defined by the successful and smart integration of naturally observed, low-latency, high-volume data with high quality survey or convenience sample data, then the ability to understand bias, or potential bias, in the data is paramount.

This paper summarized research to date on the requirements of a calibration panel to provide behavioral estimates to optimally link disparate consumer and media databases. More specifically, we looked at the case of video consumption data collected from a non-probability panel where the OTT estimates had no external benchmark data. The results of our fusion exercise showed the Demo fusion used 71% of the available SAM donors and, that the average Live rating between the fusion and Nielsen now produced a .94 correlation which was an improvement over the original correlations from either the NCS or SAM standalone data. The fact that correlations to Nielsen Live ratings went up across each fusion model speaks to the representativeness of the probability sample. Fusion models typically produce results that diverge from correlations with an independent benchmark, and despite the fusion clearly missing sources of variance with in the data, the probability sample quality of the NCS sample compensated for this.

While the Demo model was strong in terms of the correlation to topline ratings, it did not preserve internal relationships on product and brand preferences as was clearly visible in the demographic analysis of this research. The two other fusion models were better at maintaining internal relationships, but not to the extent required to capture rich consumer profiles with TV/video consumption patterns and help explain the 'why?' behind TV/video consumption by device and channel. While "split weights" fusion models are designed to better preserve internal relationships and topline currency levels, they have only been implemented in applications where both databases are strict probability samples (see Chan, Pellegrini & Whithers, 2011). This supports the need to develop calibration models that can further incorporate behavioral measures that can be used as linkage variables or within calibration techniques to build models that maintain internal consistency.

Calibration model research is resource intensive, but lays the foundation to calibrate virtually any other transactional, large-scale data set and ensure it aligns with a common set of variables that are deemed to be representative of the consumer population. While integrating disparate data sources is challenging, it offers unique opportunities to develop and evolve methodologies that uncover real insights while minimizing the measurement noise caused by integrating non-overlapping data sources.

Several key recommendations were produced by the research so far:

- Calibration panel is necessary to combine probability and non-probability data, and it must be a probability sample
- Donor-recipient matching should be used with a tool that is able to force 100% donation
- Fusions should include demographics, geographics and DriverTags™ as fusion hooks: not 50 clusters of DriverTags™ as in the pilot but the full array of 265 DriverTags™ themselves
- Continued Calibration Research might be able to identify a small set of psychographic and attitudinal variables that perform so well as fusion hooks that they should be collected on both sides of the fusion.

In the future, the ability to integrate naturally occurring and panel/survey data using best practices will mean that practitioners will be able to make decisions on one view of reality that maximizes utility and the ROI of the decisions made.

References

Chan P., Pellegrini, P.A., and Withers H. (2011) Canada's Cross Media Consumer Database: Methodology and Validation. Print and Digital Research Forum, San Francisco, USA.

DiSogra C. et al, (2011) Calibrating Non-Probability Internet Samples with Probability Samples Using Early Adopter Characteristics. JSM Proceedings. Alexandria, VA.

Harvey (2004) Better Television Audience Measurement Through the Research Integration of Set-Top Box Data. ARF Week of Audience Measurement Symposium

Kott P.S. (2017) A Partially Successful Attempt to Integrate a Web-Recrutied Cohort into an Address-Based Sample. INPS Paper.

Neigher D., Pennay D.W., Ward A.C., Lavrakas P.J. (2017). Investigation into the Use of Weighting Adjustments for Non-Probability Online Panel Samples. INPS Paper.

Palet C. and Engel W. (2016) ARF Re:Think Conference Presentation. NY, NY.

Pellegrini P.A. (2006) The Mainstream and the Clickstream. 1st Annual ARF Audience Measurement Symposium. NY, NY.

Pellegrini P.A. and Meierhoefer C. (2011) Advances in Digital Measurement: Mobile Web Usage, Bias Correction and Universe Coverage. Print and Digital Research Forum, San Francisco, USA.