ADVANCES IN DIGITAL MEASUREMENT: MOBILE WEB USAGE, BIAS CORRECTION AND UNIVERSE COVERAGE

Pat Pellegrini, Ph.D. and Cameron Meierhoefer comScore, Inc.

Introduction

This paper reports on the very latest technology and methodology for measuring the digital media landscape, and provides a view to the future of digital measurement. In particular, accounting for the complete audience of a site requires the extended projection of total unduplicated reach for Internet vehicles beyond the traditional home and work universes, including mobile and "shared access" universes. Device census data permits an understanding of the additive tonnage of Internet use across these universes. While de-duplicating reach across universes poses a unique and important challenge, there is potential to generalize findings to cross-media unduplicated reach and solutions to this are presented. Additionally, device census data can be leveraged to develop behavioral weights for bias control of a non-random selection methodology, thus eliminating the impact of response bias in 'persons' measurement. Such respondent recruitment techniques are essential to maintain the robust sample sizes required by the long tail of digital.

By embracing instead of fighting an environment of technological complexity, and bringing fresh thinking to currency development, research providers can pioneer new techniques to solve problems that were previously thought to be intractable. These include integration of mobile and internet café usage into the digital measurement total universe and development of device census calibration techniques for eliminating sample bias. This paper advocates the notion of multiple data source measurement to properly measure the digital media world of the future. Finally, the path to industry acceptance of a stable, consistent, digital measurement system is discussed, including the establishment of a probabilistic reference point for million-scale panels and a research agenda for establishing multi-data source best practices.

1. State-of-the Digital Art: Advances, lessons and remaining challenges

The growth of the internet and digital media has presented challenges and opportunities to the art and science of online audience measurement. Unlike established traditional media, like TV, with an audience measurement system of relatively small and stable set of suppliers, technologies and practices that produce accepted data on size, composition and activities of media audience, the internet is a relative 'wild west' of digital measurement approaches, companies, technological innovations for measurement, conflicting methodologies all within an environment of emerging new media concepts to measure such as Social Media, Twitter, etc. At the same time, the growing demand for increasing granularity and quality of online audience data requires unique, technology-driven but methodologically sound data collection and data processing strategies. Today, online measurement strategies must take a so-called all possible worlds approach; multiple sources, multiple panels, multiple technologies and multiple methodologies are increasingly necessary to produce metrics which are undisputed in the market. This paper presents a view of the digital measurement ecosystem that is necessary to tackle the increasingly complex demands of digital media measurement including the salient issues of mobile web usage, bias correction in digital measurement panels and total universe, or cross location and device, measurement.

At once, the internet is the most measurable and challenging medium because theoretically, digital fingerprints are left behind with every touch point, across all platforms; yet all this data overwhelms practitioners with the task of assigning interactions to persons, platforms and the proper digital property. As the 'every challenge is an opportunity' cliché goes, the digital measurement challenges are partially overcome by the opportunity that digital technology itself permits, namely, low-touch, on-device, software metering. Such low-touch methodologies serve to reduce panelist fatigue, improve compliance and keep turnover to manageable levels in situation where direct on-device measurement is desired, feasible or necessary.

Scale and Recruitment Challenges

The significant measurement challenges presented by digital media demand the continued development of 'massive scale' and innovative measurement solutions. The era of unprecedented audience fragmentation and complete on-demand media consumption requires measurement solutions that can accommodate the nearly endless media alternatives coming to market and the resulting infinite possible behaviours. The research methodology challenges are clear: how do we build measurement panels at the proper scale where that scale requires millions, not thousands, of panellists? How do we do this within reasonable cost constraints? How do we remain true to established panel methodology practices?

The logical starting point is panel recruitment. Probabilistic techniques such as 'face to face' and random digit dialling (RDD) widely used to recruit panels for traditional media audience measurement had important advantages over simpler, less-

expensive, less rigorously methodologies, namely to prevent bias and so that statistical reference points such as standard errors of estimates could be readily produced. The highly fragmented digital world which requires reporting of thousands of web sites (versus hundreds of magazines, TV stations, radio stations, newspapers, etc.) requires a recruitment scale beyond all practical and financial limits for such approaches. The solution was the move to online, non-probability based recruitment which comScore actually pioneered in the late 1990's. This technique has since become the worldwide standard for respondent recruitment for panels measuring online audiences. Such a respondent acquisition technique represents a marked deviation from traditional, offline best-practices, but is a direct response to the declining efficacy of traditional offline selection techniques (specifically, telephone-based recruitment), and the need for massive scale measurement to counter fragmentation.

There are two factors that have been systematically undermining the efficacy of telephone-based recruitment. One is the continued decline in response and cooperation rates, and the typically low response rates when asking consumers to install (and retain) software on their computers. While in theory a random design provides high quality respondent selection, in practice the actual typical yield on such a technique provides levels of non-response to raise serious questions of bias. In other words, any theoretical bias reduction to be gained from such an expensive random selection technique is effectively lost by the low response rates that such a technique will inevitably yield. Second, in many countries, the relative ubiquity of mobile phones has led to the widespread phenomenon of mobile phone-only households and, increasingly, of mobile phone-mostly households (the latter are households in which the household still retains a land line, but in which the household members use mobile phones exclusively or predominantly for making and receiving calls, and are thus effectively unreachable via land line.) Since RDD surveys typically exclude mobile phones, the spread of mobile-only and mobile-mostly households undermine the quality of the sample frame from which RDD is done. Thus, even if a panel provider could achieve a reasonable response rate using a telephone-based selection technique, the panel provider would still be confronted with the systematic exclusion of persons from mobile phone-only and mobile phone-primary households. These persons have different demographic, psychographic, and online behavioural profiles than persons in landline-reachable households.

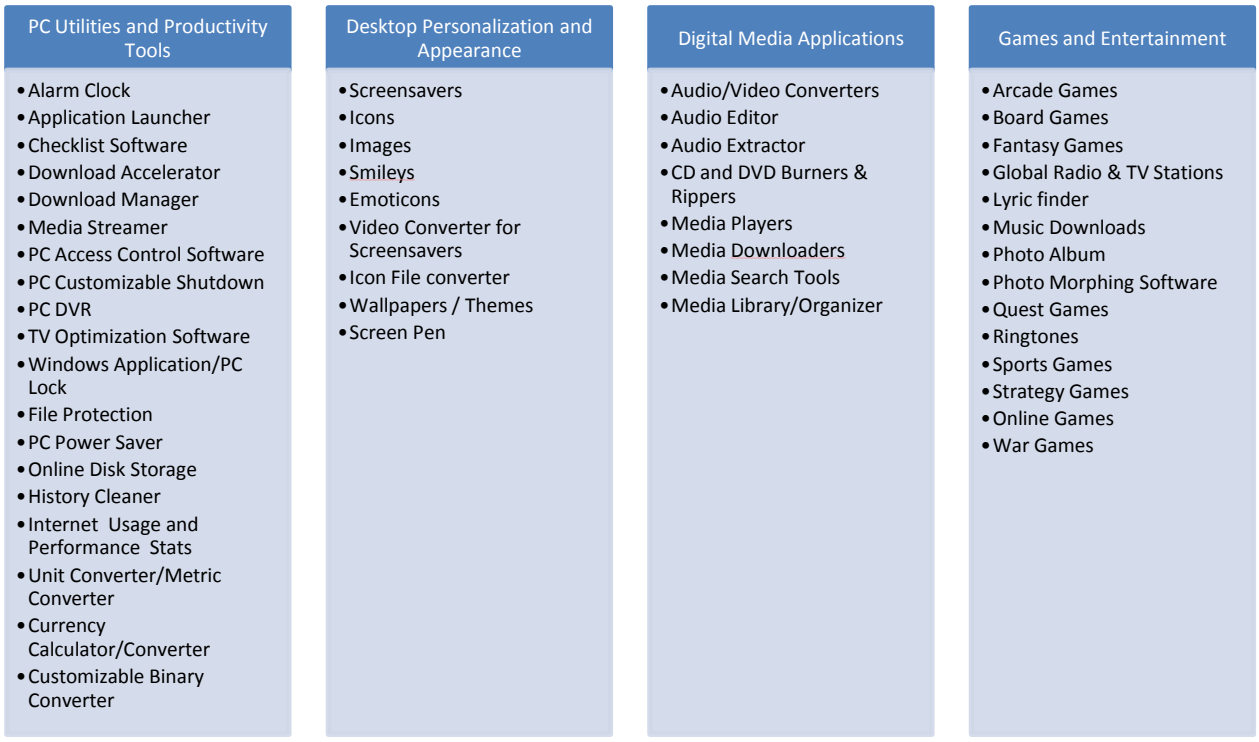
Besides overcoming the issues associated with the traditional use of landline telephone frames for RDD recruitment, online recruitment strategies permit extremely cost effective massive scale panel development. Three major online strategies are used in practice: affiliate recruiting, third-party application providers, and recruiting from within access panels (full disclosure: comScore uses the first two extensively).

The *affiliate* approach uses a network of qualified web entities where potential panelists are recruited via banner ads designed to appeal to a broad range of persons who are directed to the online intake site. Banner ads are effective for recruitment insofar as they provide an avenue for reaching a large proportion of the online population. These include individuals that might be drawn to attractive and varied offers, for example, comScore's *Trees for Knowledge*. In this special program, comScore plants a tree for each new panelist installed and so the offer appeals to panelists interested in improving the environment as well as the internet. A tree is planted on behalf of a panelist once they achieve 90 day tenure on the panel. Well over two million trees have been planted so far. It is important to note that while comScore seeks to maximize the online population that sees the offer to participate in the panel, it is also prudent to avoid affiliate sites that recruit from access panels or 'take a survey' type sites which are not representative with the online population of interest. In the same vein, it is prudent to avoid all monetary incentives since these too may skew the panel towards respondents with price-sensitive, economically driven behavior.

While a complete discussion of privacy issues in online recruiting are beyond the scope of the present paper, it is important to note here that potential panel members must be informed explicitly that they are joining an online market research panel that tracks their online browsing and perhaps other, like purchasing, behavior. For example, comScore's privacy policy is clearly stated during this stage of the recruitment process. At registration, panelists are also asked to provide demographic information, along with some other household information including household composition details. Also during registration, panelists download the software 'meter' that will then passively measure their online behavior through the device used to register them.

The third party application provider or *TAP* approach involves a partnership between comScore, for example, and various application providers who offer visitors a large and attractive array of free software, applications and utilities so long as they consider joining the comScore panel (see Figure 1 for many examples of the offerings). comScore works with dozens of such partners – over 50 in fact - who combined offer hundreds of different value propositions and recruitment campaigns. The large range of sites resulting from this ensures that a wide net is cast across the online population, one which increases the likelihood of reaching all types of potential panelists in the digital world. The ability to contact as many online people as possible with a recruitment offer is paramount in this methodology, and is consistent with the premise of a probability methodology in seeking equal probability of contact. A major advantage of the TAP approach is the advertising linkage. Many of comScore's recruitment bundles and offers are linked to advertisements by our partners across myriad sites whose reach, in many markets, is claimed to exceed 90% of the market's online population.

Figure 1: Examples of the diverse free offerings from affiliate and TAP recruitment.



A third approach that promises the scale required by digital media is *recruitment from existing panels*. Here, the aim is to recontact existing panelists from, say, an access panel, or perhaps from a large cross-section survey such as one of the high quality print surveys available. Recontact, in theory, improves co-operation rates and allows targeted recruitment in some cases. The majority of the largest panels available for recontact and recruitment are access panels and most of those have migrated online, facilitating the opportunity for PC based device measurement. However, as with the affiliate and TAP strategies discussed above, most access panels have been built via low-cost convenience sampling techniques in an effort to maximize sample size. Unfortunately, this non-probability framework does not satisfy the traditional requirements associated with media measurement, so a further challenge is how to reconcile the massive scale recruitment requirement with the probabilistic reference point required by users of audience measurement data. We return to this important question in section 3.

2. Impact of Mobile Devices and Tablets on Digital Media Methodology

In conjunction with the digital media's research challenges, the digital media ecosystem has continually evolved. Early challenges included the emergence of dynamic delivery technologies like AJAX and JSON that required measurement to evolve baseline concepts like page views, and the advent of online video has required evolution in metering technology. While these technology changes presented measurement challenges, the solutions required advancing metering technology that could be deployed within the existing measurement methodologies.

The emergence of Mobile and Tablet devices as meaningful points of consumer access to digital media presents a deeper more fundamental challenge to digital media measurement. The digital landscape is fragmenting. This fragmentation manifests itself not just in the consumer behavior we must measure, but also in the observations available to researchers, and the required methodologies.

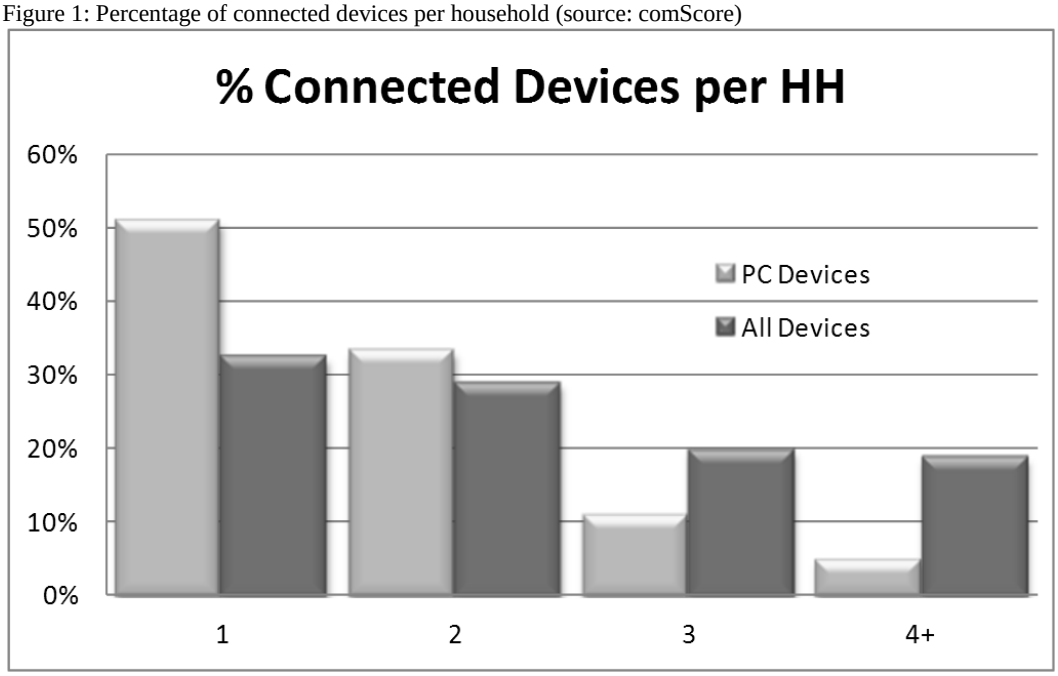
There are three dimensions of fragmentation that will impact and shape digital media measurement, and these are discussed, in turn, below.

Platform Fragmentation

Media measurement must gain visibility across a variety of technical platforms (computer, mobile handsets, and tablets). Importantly for market researchers, unlike the personal computer market where Microsoft Windows has long dominated Operating System market share, there are many competing platforms, devices and operating systems in the mobile world. In fact, in the UK, where comScore in partnership with the GSMA processes Mobile Network activity, access to digital content occurs using more than 1,000 distinct devices.

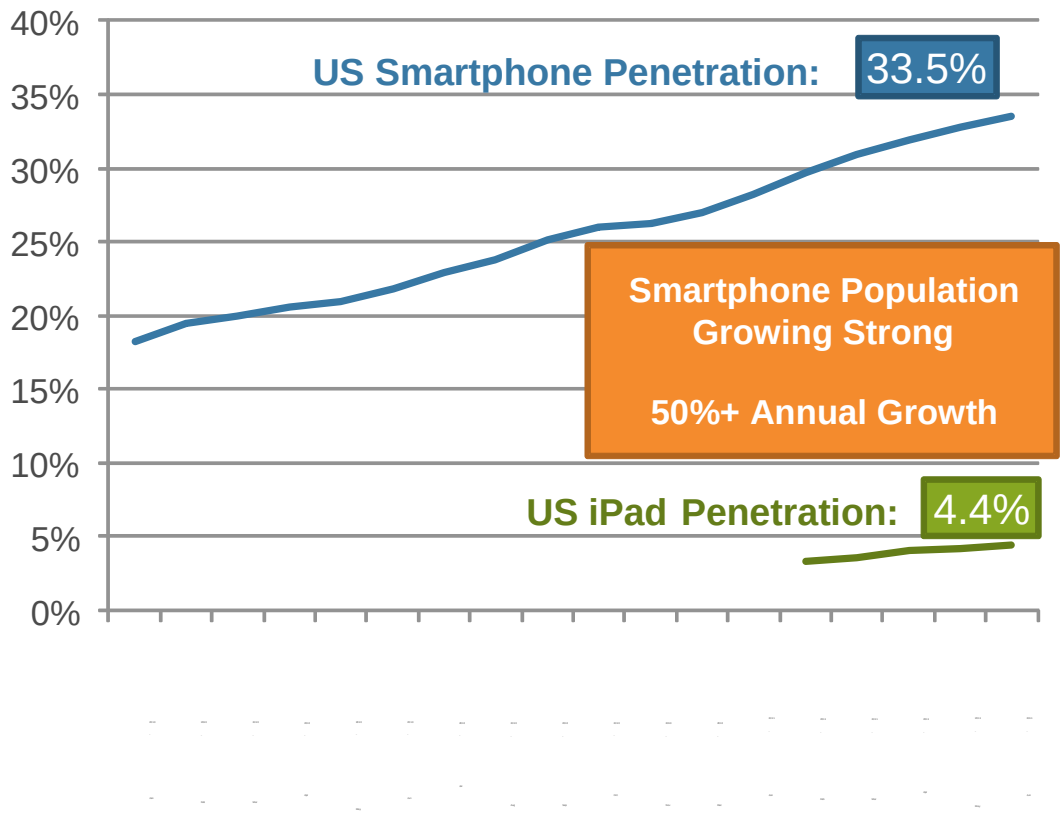
Furthermore, each distinct device may deliver a generic browser experience, but also device or operator-specific environments (including apps) that may only be available on that device. Measurement must occur across all devices to capture the full extent of behavior, and the total audience for any given publisher. The trends are undeniable. comScore recently profiled the makeup of the digital home using aggregate site-tag data. If you consider the digital landscape as if personal computers were the only device, slightly more than 50% of connected homes still have only a single PC. Once you consider the full digital landscape, including PCs, mobile devices and tablets, the profile is quite different. Fewer that one-third of connected homes have a single device, and nearly 20% have four or more connected devices (see Figure X).

Figure 1: Percentage of connected devices per household (source: comScore)



Furthermore, device fragmentation is just getting started. While the Smartphone population is growing at a rate of 50% annually, only a third of US consumers have adopted Smartphones. Tablet adoption - while the subject of much attention in the media world and in the press - remains in its infancy, with Tablets in use by than 5% of US population (see Figure 2 below). Strong secular growth is expected in both areas which will only amplify the importance of Device Fragmentation and a driving force in digital media measurement.

Figure 2: US Smartphone penetration. (Source: comScore)



Attention/Consumption Fragmentation

The emergence of Mobile and tablet platforms does not simply present a need to measure each platform properly, but also the need to measure shifting behaviors across platforms. If we could assume that people use their mobile devices exactly as they use their computers, then mobile and tablet behaviors could be directly inferred from computer-based measurement. The unfortunate reality for researchers is that this assumption is deeply flawed. People use mobile and tablet devices differently that they use their computers, and as certain digital requirements are satisfied by mobile devices or tablets traditional computer activity may diminish. Capturing the true nature of digital behaviors across platforms can only come from directly measuring how people use emerging platforms, and how that use relates to on other platforms.

Table 2 below shows the results of a recent comScore study of cross platform usage from a small subset of iPhone and iPad panelists who are members of comScore's PC panel. The results demonstrate how wildly varied incremental usage is across different types of content. Categories like Maps show significant incremental reach as many users access mapping services solely on their iPhones. High frequency categories like Social Networking show higher levels of overlap, as users are more likely to access using multiple devices, but including mobile platforms more than triples the amount for duration.

Table 1: iPhone & iPad User Incremental Usage vs. PC

Category	Incremental Reach	Incremental Duration
Total Internet	0.2%	2.0x
Maps	56.8%	9.2x
News/Information	28.9%	1.6x
Sports	26.8%	1.2x
e-mail	23.6%	1.9x
Business/Finance	21.7%	1.3x
Social Networking	12.5%	2.8x
Retail	8.7%	2.5x
Search/Navigation	6.9%	1.4x

Consumers do not shift all digital requirements to their mobile or tablet devices. They shift certain requirements, and potentially build new requirements. Media Measurement must quantify the degree of overlapping activities across the growing number of platforms an individual may choose to access content.

Measurement Fragmentation

Online Audience Measurement has long been founded on panel-based measurements built from personal computer software meters. Device fragmentation alone presents material barriers to building economically viable meter panels. The best case scenario would be that a metered-panel approach can help to measure *some* of the platforms, but not all of them. At the same time, online audience measurement has evolved to integrate site-centric measurements (typically deployed for server-based Web Analytics) to improve precision and stability of panel-based estimates (see, for example, comScore's Unified Digital Measurement™, or UDM, below in discussion of measurement ecosystem). Additionally, mobile-specific methods have emerged that leverage network-centric mobile operator logs for Audience Measurement – a robust source of measurement, but one that will not be available in every market.

Overall, digital media provides multiple potential sources for measurement. Panels, operator networks, and individual sites each have opportunities to contribute to measurement. But no single source will be able to answer the full range of market research questions at the scope, scale and depth demanded by the emerging multiplatform world. The goal is to find ways to leverage available measurement sources as compliments to build robust market research services that meet the demands of digital media. Researchers must pursue new methodological standards that focus on how best to deploy a *Measurement Ecosystem* that can systematically address any biases arising from the use of any individual data source.

Towards a Digital Measurement Ecosystem

At comScore, we have been engaged in building multiple measurement-source methodologies since 2009 when we began to integrate site-tag measurements into our panel-based audience measurement methodology, we call this Unified Digital Measurement (UDM). UDM refers to the specific technique comScore uses to combine the panel and census data sets into a seamless whole. been deployed in syndicated audience measurement systems to develop and provide the data (including cross-site duplication, reach and frequency) that buyers and sellers use to transact advertising. Site-centric, census measurement has been deployed site by site, to provide Web Analytics "internal" data for site management and optimisation. comScore believes that, when *properly executed*, a multi-source technique is the best approach to holistic digital audience measurement. Proper execution involves integration of the two data sets in a fashion that allows each to provide the things it does best (the panel, providing data on person-level behaviour; the census data, on the behaviour of the site and its servers—how many tag requests, how many cookies, etc., see Table 2 below) Then, the integration of these two data sets can yield the most accurate and actionable counts yet of audience exposure and consumption at the person-level.

Table 2: UDM as an example of multi-source approaches used to complement each other. (source: comScore)

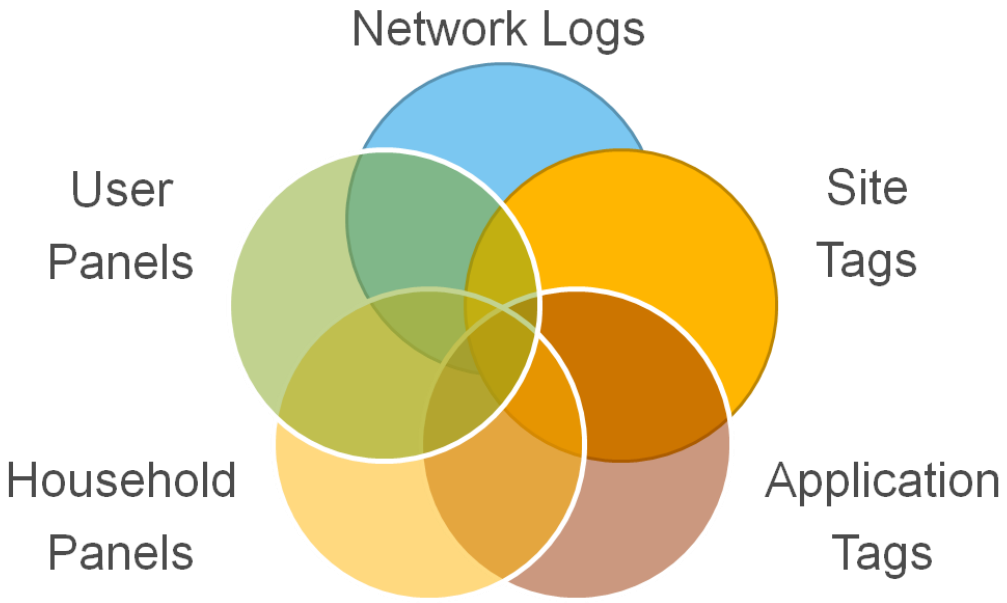
Data set	Issue	Unified Resolution
Panel	Challenges measuring the entire work universe	Census data assures that all work activity is credited
Panel	Volatility in the long tail	Census data introduces increased granularity and stability for smaller entities
Panel	Biases accruing from respondent selection	Census calibration
Census	Impossible to translate cookies to persons, except for sites with restrictive login requirements	Panel data provides cookie-to-person conversion factors at the site level
Census	Incomplete coverage of tags (e.g., tags falling off a part of the site)	Panel data allows for identifying and quantifying the extent of coverage issues

Moving forward, in every market in which such a system is economically feasible, comScore intends to offer Unified Digital measurement for both computer accessed Internet and mobile measurement.

In the constant attack on the issues of Platform Fragmentation and Attention/Consumption Fragmentation (as outlined above), we have identified five key components of the Digital Measurement Ecosystem as shown in Figure 3. Each component of the Digital Measurement Ecosystem contributes unique visibility to the digital landscape. As such, each component provides checks and balances for understanding the quality and accuracy of other components. We will review each in turn below before discussing key questions for constructing industry accepted Measurement Ecosystem methodologies.

User Panels
Traditionally recruited panel-based measurement remains the cornerstone of digital media measurement. However, difficult to measure populations such as Work and Shared Use, and other micro populations continue to present challenges to a panel-only methods. Yet, panels remain the most robust source for audience attributes such as demographics, and will continue to provide important linkages to established media measurement vehicles serving print and TV.

Figure 3: The Digital Measurement Landscape.



In a world of Device Fragmentation, the expense associated with building Panels capable of metering across the full range of connected devices and operating systems becomes prohibitive. For example, unlike the PC world of the late 90s, where a single Windows meter can measure 95% of the digital universe, there are dozens of mobile and tablet Operating Systems. Efforts to build and maintain robust metering technology is further complicated by the rapid technological change of available operating systems which require near constant reengineering to remain current.

Site-centric Tag Measurement
Site tags, traditionally used in the Web Analytics industry, provide an important source of volumes estimates in audience measurement. The primary challenge with Site tag measurement is that it does not deliver a robust unique identifier. Measurements can include unique Cookies, which are a proxy for a unique browser that is subject to a variety of acceptance and reset conditions that can lead to dramatic inflation in unique estimates if not corrected. We consider cookies to represent 'variable-quality' unique identifiers which deserve specialized treatment, but can be useful in a variety of research applications nonetheless, particularly due to the scale at which they are available. In conjunction with panel-based measurement, where both cookies and persons are available, site-tag cookies biases can be directly observed and corrected (this is the essence of UDM, described elsewhere, see Pellegrini, 2009). Site-tag measurement, at scale, can also provide important source for improvements to panel methodology. While site-tag measurement is limited to participating sites, comScore has developed methods for integrating site-centric observations to calibrate panel-based methodology that advance panel-only methods beyond demographic weighting. We return to this discussion in the next section of the paper because it is a significant and distinct topic to explore the efficacy of device census level information juxtaposed with non-probability based panel designs as discussed in this paper.

The dominant strength of Site-tag measurement is platform coverage. Properly instrumented site-tags will provide a complete view of tonnage distribution across platforms. While nuances exist with how each platform manages Cookies, and thus how representative they are of individuals, site tags will deliver accurate distributions of total requests across platforms, which is critical to managing digital fragmentation.

Network-centric Logs

Anonymous Mobile Operator logs provide a robust measurement, where it is available. Unique device identifiers are reliable, with the reasonable assumption that each device is used by an individual, and content consumption is complete, covering standard web, mobile web, and mobile app traffic. comScore has been working with the GSMA and the top four Mobile Operators in the UK, for two years, and have developed a comprehensive methods for adapting network-level measurement for audience measurement.

The main limitation of this source is that is covers only activity carried on participating mobile networks. Mobile operator logs do not include wi-fi traffic, which are typically carried by cable or fiber providers. Many smartphones connect to regular networks automatically, and up to 50% of Smartphone activity can happen over wifi connections.

The strength of Network logs is unparalleled depth and scope. Panel samples in the PC world can reach into the millions. Mobile panels struggle to break 10,000. Operator Network measurements are based on census, not samples of any kind. Issues with tough-to-recruit sub populations are eliminated. Small sample volatility issues are eliminated. Despite some significant blind spots, network-centric measurements – where available – can provide a strong foundation for digital measurement.

Dynamic Empanelment

Dynamic Empanelment is the general practice of sampling from a data source with variable-quality unique identifiers, such as those generated from site tags. These measurement sources provide census-scale measurement without durable identifiers, in general. But subpopulations within these sources are, in fact, durable, and can provide samples that can be several orders of magnitude larger than what can be raised on a traditional metered panel. These can be powerful assets in the service of Audience measurement in a fragmented digital landscape.

Three forms of Dynamic Empanelment have been deployed in either test or actual measurement. The most basic form is the Variable Lifespan Dynamic Panel (VLDP) , which is based on the 'Durable Cookies' concept established in Web Analytics that allows into analysis only those cookies that were present for the duration of the measurement period. One extension of the VLDP, is the Attributed Variable Lifespan Dynamic Panel (AVLDP), which proactively builds an attributed cookie panel through non-metered recruitment. This can be a useful practice for platforms that are difficult to meter. In these cases, it is acceptable to trade panel churn for panel sample size, assuming the economics make sense. The third variation is the Common Network Dynamic Panel (CNDP), which samples subsets of Static IP addresses. CNDP samples can be leveraged to measure cross-platform usage by consolidating activity across platforms within households (typically shared wi-fi connections). In combination with other Dynamic Panel techniques it is possible to derive relative overlap rates across publishers that can reveal platform overlap and estimates of true incremental platform audiences.

Next Generation Digital Measurement Challenges

The combination of data assets within the Digital Measurement Ecosystem empowers researchers to address the many challenges of a fragmented digital media landscape. It may provide deep insights along a number of critical research dimensions, but the industry still lacks a common understanding of best practice methods. The fundamental methodology question centers around how to estimate audience size and demographic composition across metered and non-metered platforms, while estimating the extent of overlapping usage across platforms required to calculate unduplicated audiences.

Site-tag based measurement provides insight into usage across platforms, but variable-quality unique identifiers such as cookies require correction. Panel measurements are a good source of correction where available, but researchers must also address the fact that cookies are device-dependent. Each device used by a consumer carries a different cookie. Household panels, such as a Common Network Dynamic Panel can provide observations of cross platform usage, but observations will be limited to in-home usage and have limited visibility into audience composition.

Full market measurement across sites that have chosen to tag as well as those that have not chosen to tag presents another key methodology question. Assumptions that PC behavior predict mobile and tablet behavior do not synch with reality, and provide no path to representing mobile-only or tablet-only services. While researchers can agree that it is valid to report tagged sites individually, a methodology that cannot report all media entities together falls short of market demands. The challenge to the market research community is to engage in collective assessment of Measurement Ecosystem methodologies that go beyond the traditional assessment of panel recruiting and metering methodologies. To this end, comScore is actively developing techniques that address fragmentation within the digital media landscape, as are our competitors, but the market demands assurances that methods are sound, biases are corrected and strengths and weaknesses are transparent.

3. Establishing Next Generation Best Practices for Digital Media Measurement

Researchers must continually seek to expand the technological and methodological boundaries of digital audience measurement to stay ahead of the rapidly expanding and changing digital media landscape. While much of the methodology discussed in this paper focuses on industry willingness to build and validate completely new and unique solutions, it is recognized that it is also important to bring the methodological discussion back to the point of origin where traditional and digital media collide, or, as the authors prefer, coalesce.

We have argued that the measurement of the digital world requires a 'measurement ecosystem concept' and must look beyond a single measurement source, methodology or technology so as to capture the full digital media behavioral landscape. In fact, this 'all possible worlds' notion is central to future-proofing comScore's digital measurement strategy. Of course, this is not novel as even the traditional analogue TV measurement combined paper diaries and electronic meters in the same markets.

At its core, the measurement ecosystem contains concepts such as 'user panels' which remain the cornerstone of media measurement primarily because they provide important person based demographics, durations and behavioral data used in media planning and buying. In addition, site-centric tag measurement, network centric logs, and dynamic empanelment, together with user panels, form the wider components of this ecosystem and are the basis for continued research and development. As leaders in the development of methodologies to measure digital media across a growing number of platforms, comScore is aware of the challenges involved in the validation of these methodologies both individually and as a composite system.

Establishing a Probabilistic Reference Point for Massive-Scale Panels

The user panel recruitment process, as presented in section 1, is based on a non-probability online sampling methodology. To understand the implications of this approach from a validation standpoint, it is useful to view statistical sampling more generally, and contrast probability and non-probability approaches. Researchers collect information by a wide variety of methods ranging from experimental designs in the physical sciences through to more common social science survey techniques, and both probability and non-probability approaches are used.

A probability based sampling methodology starts with a population of interest, for example, the online population from home or work, and then a sampling frame is established. If the population of interest was all persons aged 21 year or over in a private household, the frame might be geographically delineated and consist of a listing of all private dwellings in that area, including all the people who live in them. Clearly, such a frame is unavailable for the online population in any market, but we return to this later.

A probabilistic algorithm is used to select a sample from the frame where the key is that every element of the frame has a known chance of being selected, and that we can calculate the probability of selecting the sample we end up with. The probabilistic algorithm specifies who is to be interviewed, surveyed or recruited to join the sample; no judgment is involved. The overall aim is that, if the sample were to be repeated many times, the expected value of the results from the repeated samples would be the same as the result we would get if it were possible to survey the whole population.

An advantage of probability sampling is that it is possible to calculate a sampling error for the results in a straight-forward manner where the sampling error measures the amount of variation between the results due to the sampling alone (not to be confused with measurement error). It is a measure of the quality of the sample design, and of the results. Often referred to interchangeably as 'reliability' or 'relative stability,' sampling error provides a mathematical way of expressing the stability of estimates from samples of particular sizes.

While a probability sample is typically viewed as a prerequisite for developing accurate estimates of the amount of sampling error associated with survey based estimates, sample sizes and a host of other factors may influence the true reliability of a real-world sample. For example, in traditional and digital media measurement platforms, such as an internet or TV measurement panel:

- Selection may be in multiple stages – and involve use of establishment surveys to ensure representativeness of the sample with respect to a range of relevant factors
- Samples may be clustered - all members of a household are included although co-operation may have been obtained froma single individual
- Samples may be stratified - geographically, and by other variables including cost and expected response rate

- Data are weighted, sometimes in complex ways - this adds to the variance (sampling error) even though it is useful to reduce potential bias
- Adjustments are made to the data - visitor viewing, live versus playback, etc.
- Complexity of various estimates themselves - simple ratings (proportions) to shares each have their own error formula
- Repeated measurement - use of averages over reported time periods (days to weeks) impacts standard error

While for online digital media measurement specifically:

- Online recruitment partners may vary month to month and from each other
- Recruitment campaigns may be staggered across time periods and of varying sizes in terms of their reach
- Offers in exchange for participation may change vary the propensity to participate
- Location of use such as Home, Work, Shared or Mobile impacts

All of these factors combine to make real-world samples behave differently than a 'textbook' reliability formula based on the so called simple random sample (SRS) sample design. Recall that SRS includes the requirement that all elements of the frame have a known and independent probability of inclusion in the sample. While perfectly legitimate statistical sampling practices like stratification and multi-stage sampling alter this assumption, and can be dealt with via well-known extensions to the basic standard error formula, many of the deviations from SRS in both traditional and digital media measurement cannot. So, as complexity is added to the sample design, even more complexity is added to the formulae for calculation of the standard errors.

These differences are often summarized and captured in the measure called "statistical efficiency" (or sometimes called the design effect, DEFF) which describes whether a particular estimate has a real standard error that is larger or smaller than SRS theory would estimate. So, a statistical efficiency of 1.25 indicates that the rating in question has reliability equal to an SRS sample that is 25% larger, whereas efficiencies of less than 1 indicate that a particular sample is less reliable than an SRS sample of the same size. Said another way, the DEFF measures the extent to which the actual variance deviates from the SRS variance of an equivalent sample size. DEFF is related to "effective sample size" simply as n/DEF where n is the actual sample size. If the actual sample size is 12000, but the DEFF is 2 then the effective sample size is 6000. That is, estimates derived from a survey with the given sample size have the same precision (predictive power) as a result from an SRS based survey of 6000.

So we can see that probability samples in media (and other social and physical science) research require levels of complexity and assumptions that stray from the SRS assumptions and, in turn, look to computer intensive methods such as bootstrapping and jackknife replication to calculate DEFF to determine standard errors. Non-probability samples carry the burden of having to prove they are largely free of subjective sample selection bias, while not having comparable estimates of precision to probability samples. In this way, both probability and non-probability sampling run into similar validity concerns.

Take a non-probability approach like quota sampling as an example: the key idea is to match certain characteristics of the survey sample with the population (in terms of age and sex say) by filling quotas for each of these characteristics. The assumption is that if the sample matches the population on these characteristics, it may also match the target population on the quantities we are trying to measure (say, digital media consumption). While at a glance this approach resembles probability based stratified sampling, it differs significantly because interviewers have some element of judgment in their probability of selection. That is, interviewers may consciously or subconsciously choose non-threatening or easy-to-approach respondents, or those who are more easily contacted. A distinguishing feature of a probabilistic sample is that interviewers would have no choice about who they are to interview.

The inherent need to judge the quality of a judgment sample (or non-probability sample) leads to the same need for the derivation of an empirical distribution function (EDF). Besides the complexity of probability samples pointing to non-parametric inference techniques like the jackknife, the declining response rates associated with surveys in general mean that probability designs, no matter how complex, also suffer from validity concerns related to sample selection bias. Here we can go back to the issue of having a sampling frame of all online persons – since these do not exist; probability and quota sampling are inherently difficult for online measurement unless recruited from a previously recruited and classified master sample.

Jackknife Replicates, Probability and Validation

A fundamental task of quantitative media research is to make probability-based inferences about a population characteristic based on an estimator using a sample drawn from that population. The discussion above has shown that increasingly complex probability-based sample designs for digital media preclude the use of traditional parametric inference (like the central limit theorem). In fact, both probability and non-probability based sampling require non-parametric inferential techniques to examine the reliability of sample statistics. The inherent advantage of probability samples is lost in the complex world of

digital measurement[1], while the apparent disadvantage of non-probability samples can be overcome so long as full spectrum validation is available through state of the art methodology like the jackknife and probability validation panel (PVP).

The 'jackknife' is a computationally intensive, non-parametric technique to assess the variability of a statistic by examining the variation within the sample data, rather than through the use of parametric assumptions, and is widely used for making such inferences across complex sample designs[2]. It differs from traditional parametric approaches to inference in that it employs large numbers of repetitive computations (or 'resampling') to estimate the shape of a statistic's sampling distribution rather than making strong distributional assumptions and using complex analytic formulas. This allows media researchers to make inferences in cases where such analytic solutions are unavailable or untenable. The jackknife is not a statistic *per se*. Rather, it is an approach to using statistics to make inferences about population parameters that relies on an analogy between the sample and the population from which the sample was drawn. The central idea is that it may sometimes be better to draw conclusions about the characteristics of a population strictly from the sample at hand, rather than by making perhaps unrealistic assumptions about that population.

The computational exercise of producing jackknife replicates permits a comparison of traditional standard error versus that computed from the large number of jackknife replicates which produces, mathematically, an estimate of the DEFF and, in turn, the statistical efficiency levels of the sample. Therefore we can rigorously determine the deviation of estimates from non-probability panels over time given the fluctuation in sample sources, for example, as well as disentangle efficiency losses attributable to weighting. Furthermore, generalized error estimates and confidence intervals for key measures are obtainable.[3]

While undoubtedly valuable, the jackknife technique only completes part of the validation of non-probability online sampling and recruitment methodology. The reason is that the sample derived EDF is assumed to be a good estimator of the population distribution function (PDF) that generated the sample in the first place. This means we believe that a representative sample of all the possible distinct values of the population is found in our data. This is problematic where sample size is small and so less likely that all of the relevant characteristics and behavior of the population will be represented. This is another reason why relatively large user panels will remain an issue for digital media measurement in the future. In the end, even a technique which minimizes *a priori* assumptions still faces the question of whether the EDF is a good estimator of the PDF in situations of complex non-random sampling.

For this reason, probabilistic validation panels (PVP) recruited using either two-stage (establishment survey followed by recruitment) sampling or area probability (one stage) sampling are necessary to further compare the key metrics, effective sample size and any bias the online non-probability panel may have[4]. While a probability panel large enough to permit highly granular analysis is prohibitively expensive, and largely unattainable given prevailing response rates via phone, internet or face to face recruitment, a sample of 2000 persons with full demographic information should provide substantial learning in most mature online markets.

The PVP can be used to provide baseline demographic and behavioral profiles at topline levels for larger sites to compare with those from online recruited user panels. In addition, the probabilistic nature of the panel permits an evaluation of the effective sample size, across each demographic segment, of the online panel. This provides another view of the effective sample size versus the jackknife replicate analysis discussed above. The PVP will also provide a view of any sample bias and suggest corrective weighting at the demographic level. Essentially, the combination of jackknife non-parametric estimates along with a parametric PVP approach triangulates on the reliability and bias questions associated with best practice online measurement methodology with the goal of minimizing required assumptions and providing the least biased minimum variance estimators.

The notion of a PVP is significant for the advanced validation of device census calibration techniques as well. Recall from the earlier discussion of site-centric tag measurement that the natural distribution of usage intensity may be used to viewed through site-centric measurement at scale. To the extent that usage intensity is a proxy for online behavior more generally, these metrics can be integrated into panel-based methods to move beyond basic demographic weighting, and thus the non-

[1] Many statisticians would argue that most surveys of human populations do not meet and cannot satisfy the criteria of a probability sample as the size and nature of the selected elements that cannot be contacted and/or will not agree to allow themselves to be surveyed is such that the operational device of claiming that they form a random subsample of the selected sample is clearly not sustainable. A similar argument is that the probability of a randomly selected element completing the survey is unknown and for some elements may be zero, so that classic inferential conditions cannot be met.

[2] The Jackknife is closely related to the more general Bootstrap technique; the latter uses sampling with replacement while the former uses the 'drop one out' approach.

[3] The estimated error derived from using a jackknife method measures the dispersion of sample means around the corresponding average mean; however, with non-probability sampling the average mean is not necessarily the same as the population mean. In other words, the jackknife estimate does not eliminate bias in the estimates resulting from a non-probability based sample. Survey statisticians would now argue that other information has to be found to either confirm that the sample estimate is unbiased or suggest factors that can be applied to remove such bias that does exist.

[4] Area probability sampling has the advantage of single stage recruitment that limits the opportunities for equivocation about joining a panel and elegantly overcomes the cell-only and cell-primary limitations of random digit dialing (RDD) approaches. The two stage establishment survey/recruitment approach does not require extensive granular block listing of area probability and more elegantly deals with substitution for refusals, but may be hampered by RDD challenges.

probabilistic nature of the panel has no bearing on its usage. However, the important assumption underlying site-centric or device census calibration is that the underlying user panels are broadly representative in terms of cookie deletion since we view census behavior through the lens of panelist cookie consumption. This means that validity of any device census information is tied to the extent to which panelist cookie consumption is representative of universe cookie consumption, and this can be judged by comparing cookies per person across categories between the PVP and the online panel.

An equally important reason for the development of PVPs in next generation online methodology is the familiarity of this probabilistic approach. comScore has found that validation using familiar methodological territory is the ideal approach to understanding, validating and refining new methodologies for digital measurement. Besides familiarity and validation, the testing of new concepts is readily available in this environment.

Statisticians are not as unanimous as perhaps they once were on the foundations of sampling, and the past two decades has seen substantial debate in the academic and applied literature about the role of various assumptions, models and randomization in sample design. Certainly, non-probability approaches are the norm in studies of the urban population, sample surveys of businesses and even consumer price index studies. The point is that while non-probability methods may be less familiar, they do not necessarily compromise quality and rigor of survey estimates to the extent that they are not commercially valuable. Indeed, they often provide a uniquely cost-effective result about which reliable margins of error can be applied.

The theme that runs through this discussion is of 'checks and balances' - validation and comparison, assessing the rationality of assumptions, and continually improving digital audience measurement. Perhaps we can be accused of overstating the case of validating online measurement, but paramount in the advancement of digital media, especially with the dynamic and challenging ecosystem of measurement publishers and agencies currently face, is a process of layered, rigorous, and digestible validation using best practices.

4. Summary and Discussion

Digital media has always pushed boundaries of established media measurement methodology and this paper focused on the very latest advances in digital measurement with a view to the future. Recent advances in mobile and tablet technology present a new set of challenges, both technical and methodological, while massive fragmentation is set to define digital media going forward. We argue that both technology based measurement and research methods must adapt accordingly. Researchers must reconcile outstanding methodological questions in order to address this new wave of challenges and we have provided examples of best practices for accomplishing this task.

While presenting the current state-of-the-art, we highlighted some open questions such as the ramifications of multi-platform measurement and the need to incorporate multiple data sources to properly measure digital media. On the former, site tags provide information across different platforms and different locations using operating system information and associated IP addresses. The main problem of measuring cross-platform and multi-location usage is the duplication of individuals with multiple access points. General data on how the consumption behavior of digital content from different platforms and locations for the different types of individuals exist, but comScore believes it is necessary to provide a more granular approach to these phenomena to take into account possible differences between these patterns of duplication for different content. For example, what is the duplication between locations and access platforms for an online newspaper against an e-commerce site? And how does that differ across demographic groups? Such information must be collected at the site level and across locations and platforms. Industry accepted principles must be established that recognize that as digital requirements shift across devices, it is critical to measure each platform directly. Usage patterns from one platform do not predict usage patterns on another platform.

The path to establishing industry accepted digital measurement methodology lies in the establishment of a probabilistic reference point for massive-scale panels and multi-source measurement. Separately, a research agenda for establishing multi-data source best practices needs to be at the forefront as the measurement ecosystem expands. For example, the potential contributions of site centric measurement are significant, but it is important to note that cookies are to be considered "variable quality unique identifiers" and require any potential or actual cookie biases to be corrected via direct observation, say, via user panel. Finally, industry participation is vital. Today, tagging is very low cost and extremely fast. Participation in tagging protocols means greater precision in measurement, the potential for cross-location, cross-platform measurement, and, in time, a standard digital measurement currency.

References

P A Pellegrini (2009) "Towards Panel-Centric Hybrid Measurement: Enhancing Online Audience Measurement" *Proceedings of the Worldwide Readership Research Symposium,* Valencia, Spain.