# People First: A User-Centric Hybrid Online Audience Measurement Model

by John Brauer, Manager,
Product Leadership,
Nielsen

## Overview

As with any media measurement methodology, hybrid online audience measurement calculations are only as good as the input data quality. Hybrid online audience measurement goes beyond a single metric to forecast online audiences more accurately by drawing on multiple data sources. The open question for online publishers and advertisers is, 'Which hybrid audience measurement model provides the most accurate measure?'

- User-centric online hybrid models rely exclusively on high quality panel data for audience measurement calculations, then project activity to all sites informed by consumer behavior patterns on tagged sites
- Site-centric online hybrid models depend on cookies for audience calculations, subject to the inherent weaknesses of cookie data that cannot tie directly to unique individuals and their valuable demographic information
- Both hybrid models assume that panel members are representative of the universe being measured and comprise a base large enough to provide statistical reliability, measured in a way that accurately captures behavior
- The site-centric hybrid model suffers from the issue of inaccurately capturing unique users as a result of cookie deletion and related activities that can result in audience measurement estimates different than the actual number
- Rapid uptake of new online access devices further exacerbates site-centric cookie data tracking issues (in some cases overstating unique browsers by a factor of five) by increasing the likelihood of one individual representing multiple cookies

Only the user-centric hybrid model enables a fair comparison of audiences across all web sites, avoids the pitfall of counting multiple cookies versus unique users, and employs a framework that anticipates future changes such as additional data sets and access devices.

nielsen

## Source Effect: Cookies, Panels or Both?

Consumer behavior serves as the bedrock of audience research. Traditionally, online browsing behavior was measured using one of two approaches. In the panel-based method, behavior of an audience sample was extrapolated to the larger population. In the site-based census model, measurement tags attached to web pages enabled census-level behavior measurement.

The most recent developments in online audience measurement leverage hybrid methodologies that combine aspects of both the panel-based and census approaches, offsetting the shortcomings of each model. The differences between hybrid online audience measurement models reside in how the data inputs are used—specifically, the relationship between the monthly audience and the monthly cookie counts.

In genetics, the theory of hybrid vigor proposes that hybrids will deliver improved performance results over that of purebreds. The same theory applies to online audience measurement. Hybrid models that go beyond a single metric forecast online audiences more accurately by drawing on multiple data sources. The goal of the hybrid audience measurement method is to project audiences based on all individuals with access to the Internet. Access points include home and work PCs, public locations (libraries, cafés), dormitories, Internet-enabled mobile devices and other devices with Internet browsing capabilities (e.g. Xbox 360). The process involves three basic steps: establishing the demographic profile of the population with Internet access, estimating the likely audience for each site, and selecting panelists to represent audience from other locations.

## Focus on Panel

Panel-based measurement utilizes a sample of online users to project to the broader population. Panel quality is critical to the integrity of the data used in audience projections.

In its "Guide to Understanding Online Measurement Alternatives,"[i] the Media Rating Council articulates the benefits of panel-based methods, namely, the robust demographic information of panels and the comprehensive online behavior of sampled individuals (as opposed to single site or campaign measurement).

The Guide also cites three critical requirements for effective panel-based measurement:[ii]

1. Participants must represent the universe being measured in terms of the relevant behaviors
2. The number of participants needs to be large enough to provide the reliability and stability required for measurement applications, and
3. People must be measured in a way that accurately represents their behavior.

Panel quality directly impacts the quality of data generated using that panel. Hybrid measurement solutions address some of the challenges associated with panel-only methods by incorporating census-level behavior. Regardless of which hybrid online audience measurement approach is selected, audience calculations are only as good as the input data quality.

Reliable audience measurement demands an ongoing investment in maintaining a high quality panel and the ability to draw on use-based observations to project activity across all sites, informed by measurement-tagged input.

## Focus on Census

Census-based models collect data via a measurement tag, a small piece of computer code appended to each page of a site, enabling browser behavior tracking within the site. When loading this code, the browser passes along basic information such as URL, operating system and browser type to the measurement firm's servers.

This provides near census-level accounting of site activity, a direct measure of consumer behavior on a site, helping understand the amount of content consumed. Once filtered for non-relevant traffic such as search bots, spiders, auto-refresh, international and similar traffic, the volumetric data provides valuable input to the models.

> *"The threshold of measurement difficulty for achieving this measure (individual site users) in a census-based environment is quite high."*
>
> –IAB, 2009

Measurement tags accurately measure site activity through an exchange of a browser cookie. Cookies identify the unique browser accessing content on a site, and are typically used to customize content and advertising messages, maintain user preferences and generate web analytics. Among hybrid solutions, only site-centric hybrid methods are dependent on cookies for audience calculations.

# One Horse or One Race?

Horse racing provides a helpful analogy for understanding the difference between panel-based and cookie-based data. Whether judging a race or evaluating a media plan, it all centers around accurate measurement.

In media planning, performance evaluation across competitors is key. In horse racing, when the playing field is level and the conditions are consistent, each race provides real-time performance evaluation of all the horses in the race. Panel data compares to a single race where all horses are running at the same time, on the same length track under identical turf and weather conditions on a consistent, comparable basis. Cookie [census] data compares to multiple one-horse races, where all the horses are running at different times, on different length tracks, and under different turf and weather conditions, so rendering data comparisons is difficult at best.

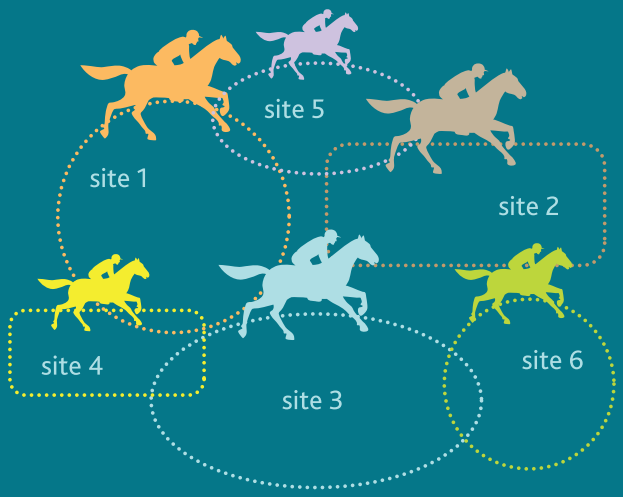
The problem comes down to this: one person equals many cookies. The cookie approach can over- or under-count due to the way people use the Web. Users access from multiple devices, share web-enabled devices and switch to browsers that may clear or reject cookies, which thereby overwhelm systems.

Cookies are demographically "blind," and unable to link surfing behavior back to audience demographics, which fails to meet even the minimal audience reach criteria established by the IAB: "…the measurement organization must utilize in its identification and attribution processes underlying data that is, at least in a reasonable proportion, attributed directly to a person."

**User-Centric Hybrid Audience Measurement**


*Consistent conditions create a level playing field and a fair race for all participants, creating an ability to effectively compare horses since they are running the same race on the same track*

**Cookie-Dependent Measurement**




*Independent conditions lead to individual races with individual results: no ability to compare horses when they are running their own races*



nielsen

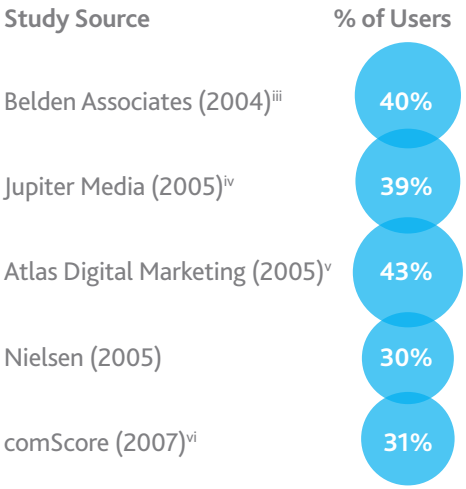
# Cookie Cutters

Despite the benefits of web analytics and user experience customization, online researchers have discovered a fundamental problem with cookies. Transient elements by design, cookies are not permanent elements associated with a single user. As a result, users who delete cookies may have many cookies assigned to them while online, resulting in inflated audience numbers.

It is this aspect of cookies that defines an underlying difference between user-centric and site-centric hybrid models. Numerous studies underscore the tenuous accuracy of cookies, revealing that cookie deletion rates range from 30 to 43 percent.

## Chart 1: Cookie Deletion Rates

***One or More Times per Month***




To further illustrate the concerns related to cookie deletion in site-centric audience measurement models, see "Estimating the Effects of Cookie Deletion" on page 5.

Even more potentially damaging to cookie-dependent models than the over- and under-counting associated with cookie deletion and suppression are:

- Private browsing
- Inherent browser bias
- Opt-in privacy legislation
- Internet device proliferation
- Cookie bloat

Private browsing is becoming an increasingly available option on popular browsers such as Microsoft Internet Explorer®, Google Chrome, Apple Safari and Mozilla Firefox. Private browsing eliminates any trace of the browsing or search history by wiping out all cookies and cached files retrieved during the online session.

Research from Stanford tested the impact of private browsing on ad tracking for three ad campaigns. Private browsing activity was detected in approximately seven percent of the cases, meaning site visitation numbers based on cookies were inaccurately under-recording unique users for these campaigns.

Further compounding the issue, private browsing utilization rates differed significantly from browser to browser, with higher adoption rates registered for Safari than for Internet Explorer. This introduces the issue of browser bias into the private browsing discussion.

Another complicating factor for researchers in the European Union will be the different ways member states choose to implement Directive, 2009/136/EC, which requires an explicit "opt-in" consent to place or read cookies.

Perhaps the most significant issue of all affecting the accuracy of cookie counts is the explosion in the number of Internet access devices. Tablets, smartphones, video game consoles and televisions now come equipped with Internet browsing capabilities, often usurping standard issue laptops and desktops as the access portal of choice.

A more connected population increases the likelihood of one individual representing many cookies. To demonstrate the phenomenon of cookie bloat, look at the case of Australia, where the number of unique browsers recorded by cookies in April 2011 was five times the number of Australians with Internet access—a clear impossibility.

# The People Factor

The potential for inaccurate reporting of unique visitors through cookie counting alone prompted the Interactive Advertising Bureau (IAB) to write into its 2009 online audience research guidelines a recommendation that audience/visitor/user data for any site "must include a component that is directly attributable to people," such as panel data.

Adjusting cookie counts using panel behavioral data may help site-centric approaches to meet this minimum threshold standard. However, the number of challenges (e.g., cookie suppression, deletion, private browsing, opt-in requirements, devices per user, users per device, browsers per device and other variables) affecting the accuracy of individual site user counts generated by cookies heightens the risk of error. This fact was underscored by the IAB, which noted in its guidelines that "the threshold of measurement difficulty for achieving this (audience) measure in a census-based environment is quite high."[ix]

# Estimating the Effects of Cookie Deletion

For illustration purposes, imagine a website with 10,000 visitors per month. Approximately 10 percent of these 10,000 people delete their cookies every day; another 20 percent delete their cookies once a week, and the rest don't delete their cookies at all.

In a hypothetical 30 day month, the following cookie counts were observed per type of visitor:


Now, calculate the audience estimated based on the cookies counted. Assuming that 20 percent of the 10,000 site visitors go to the website daily, 30 percent visit weekly, and the rest only visit once monthly, these calculations generate the following numbers:

| | Daily deleters 10% | Weekly deleters 20% | Non-deleters 70% | Cookie count |
|---|---|---|---|---|
| Visit every day (2000) | 2000 X 10% X 30 = 6000 | 2000 X 20% X 4 = 1600 | 2000 X 70% X 1 = 1400 | 9000 |
| Visit once/week (3000) | 3000 X 10% X 4 = 1200 | 3000 X 20% X 4 = 2400 | 3000 X 70% X 1 = 2100 | 5700 |
| Visit once/month (5000) | 5000 X 10% X 1 = 500 | 5000 X 20% X 1 = 1000 | 5000 X 70% X 1 = 3500 | 5000 |
| | | | **TOTAL** | **19,700** |

This highly simplified example underscores how the 10,000 individual visitors were counted incorrectly as 19,700 visitors, an audience overstatement by a factor almost twice the correct number as a direct result of cookie deletion.



nielsen

## Model Evolution

Despite pioneering the site-centric, cookie-driven hybrid methodology in 2005, Nielsen later reversed its opinion based on four subsequent years of intensive study. After vetting the approach in a number of markets, Nielsen withdrew its support formally in a 2009 paper presented to the Advertising Research Foundation. The paper cites three primary areas of concern about site-centric methodologies:

- Data is only generated for the portion of the market willing to tag
- A lack of respondent-level data sources limits the ability to drill into reports
- Demographics are not directly linked to the cookie-derived audience

These factors led Nielsen to conclude that data from the site-centric hybrid method has "limited utility for media planning because it represents a report on values and not actual people."[x]

## Assumption Proliferation

Additional research reinforced that, to achieve a truly accurate audience measurement, any site-centric hybrid online audience model must fulfill three important and distinct assumptions:

1. Site visitation and behavior on one metered device can predict visitation and behavior on another device
2. Site visitation and behavior on one metered device can predict cookie duplication on the same device
3. Site visitation and behavior on one metered device can predict cookie duplication on another device

Mathematically, the more assumptions underlying a model, the higher the probability that such a model will fail or generate inaccurate results. Given the site-centric approach dependency on cookies, it is virtually impossible to meet all three assumptions. Further confounding site-centric measurements are the mixed methodologies used to determine audiences on tagged and untagged sites.

## Cookies: No Common Standard

The problem with cookie-dependent audience measurement is the same problem all Web analytics share: each site captures and cleanses data differently, making it virtually impossible to establish any consistent benchmarks to compare one site to another. This proves especially problematic for advertisers, whose expectations are grounded in experience with reliable, robust syndicated data that fosters industry-wide comparability. Only a user-centric online hybrid audience measurement methodology delivers the uniform, industry-wide perspective required for critical media buying decisions.

As a result of cookie-based model limitations, Nielsen developed an alternative, user-centric hybrid approach.

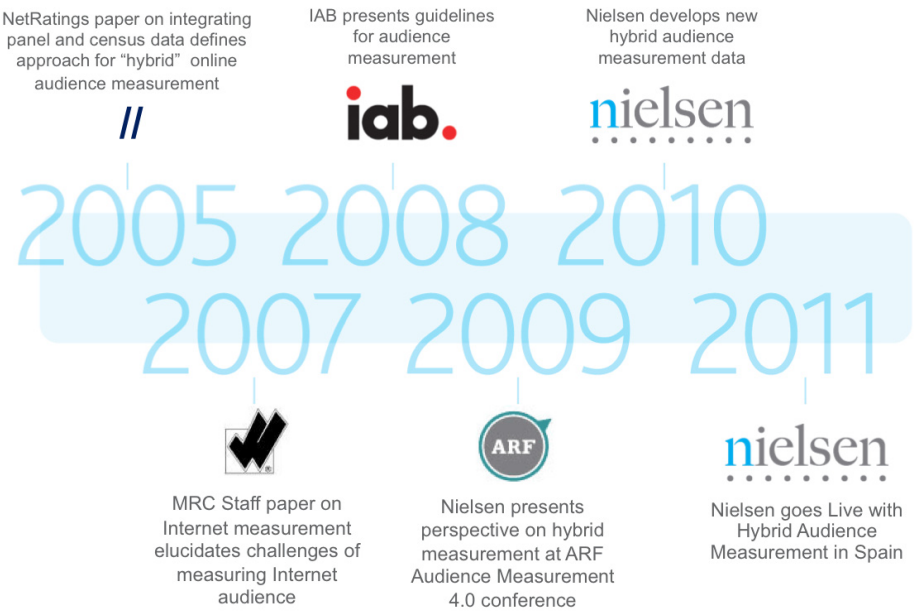
This user-centric hybrid model focuses on quality behavioral data available from both data sources—panel and census-level. This more robust, people-focused model demonstrates a greater degree of flexibility and an open design to meet future challenges such as reporting on a device- or location-specific basis and accepting input from new metering devices and new data sources that may be developed.

## Testing Assumptions

As a proof of concept for the user-centric hybrid audience measurement model, Nielsen tested the underlying premise that the online behavior of a demographic group on one device can predict the behavior of others in the same demographic group on another device (assumption #1 above).

By extension, this principle suggests that browsing behavior within the measured sample of home and work computers, when combined with census-level behavior from tagged sites, can be used to understand browsing behavior from other devices and other locations.

### Chart 2: Evolution of Hybrid Methodology



The soundness of this assumption was tested and authenticated by performing a study with Nielsen measured samples using negative binominal distribution [NBD]. The NBD technique offers a statistical framework to describe the frequency distribution of behavior, such as the percentage of homes that purchase a product and the average number of items purchased in a period, or the number of people exposed to a medium and the distribution of number of exposures to the medium in a period.

The NBD model is very well established as a means of estimating media reach and has been validated by several studies over the years. NBD enables the user-centric hybrid model to understand the totality of online access, accounting even for locations and devices other than those directly measured.

Home and work panelists that visited a selection of tagged sites were grouped into location-specific demographic and behavior groups such as: M25-45 + Heavy usage or F25-45 + Light usage.

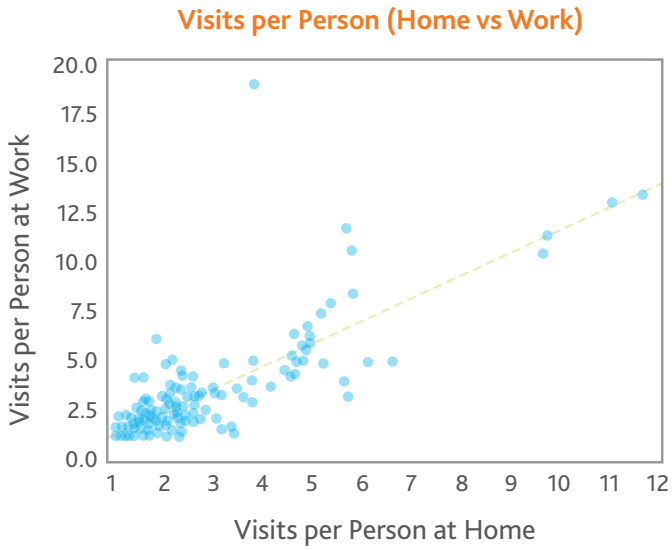
As both Chart 3 and Table 1 show, usage patterns of a demographic group with one device and location did indeed serve as a significant predictor of the same demographic group's usage on another device and location. The exceedingly low p-value and robust adjusted R-squared coefficient reinforce the strength of the finding.

Working from this established principle, Nielsen developed a sophisticated method for analyzing behavior in metered and tag-based data and generating a database that relates each action on a site with a specific individual and their demographic profile.

Unlike site-centric hybrid approaches that do not link demographics to behavior, the Nielsen user-centric hybrid audience measurement database establishes that direct link and includes all the reporting capabilities of traditional panel-based data.

Having established that user behavior on a site can serve as a predictor of cookie duplication on the same machine, one question remains: does the behavior of one individual on one metered device predict the additional cookies that may be acquired on additional devices used to access a site? To test this assumption, usage patterns for the age/sex home sample were observed to determine whether they could predict usage patterns for the same demographic strata in the Nielsen work sample. Results confirmed the correlation.

## Chart 3: One Device/One Location Predictive Power




## Table 1: Predicting Visits in One Location Based on Another Location

| Home/ Work model | Dependent Variable | Intercept | Visits/ person at home | F-statistic | p-value | Error DF | Adjusted R-squared |
|---|---|---|---|---|---|---|---|
| coefficients | visits per person at work | (0.07) | 1.15 | 315.89 | <.0001 | 182 | 0.635 |
| T-statistic | visits per person at work | (0.36) | 17.77 | | | | |
| p-hyphen-value | visits per person at work | 0.72 | <.0001 | | | | |

## Table 2: Predicting Cookie Duplication in Location Based on Another Location

| Home/ Work model | Dependent Variable | Intercept | Visits/ person at home | F-statistic | p-value | Error DF | Adjusted R-squared |
|---|---|---|---|---|---|---|---|
| coefficients | cookies per person at work | (0.12) | 1.11 | 164.2 | <.0001 | 113 | 0.589 |
| T-statistic | cookies per person at work | (1.21) | 12.81 | | | | |
| p-value | cookies per person at work | 0.23 | <.0001 | | | | |



nielsen

## Site Uncertainty

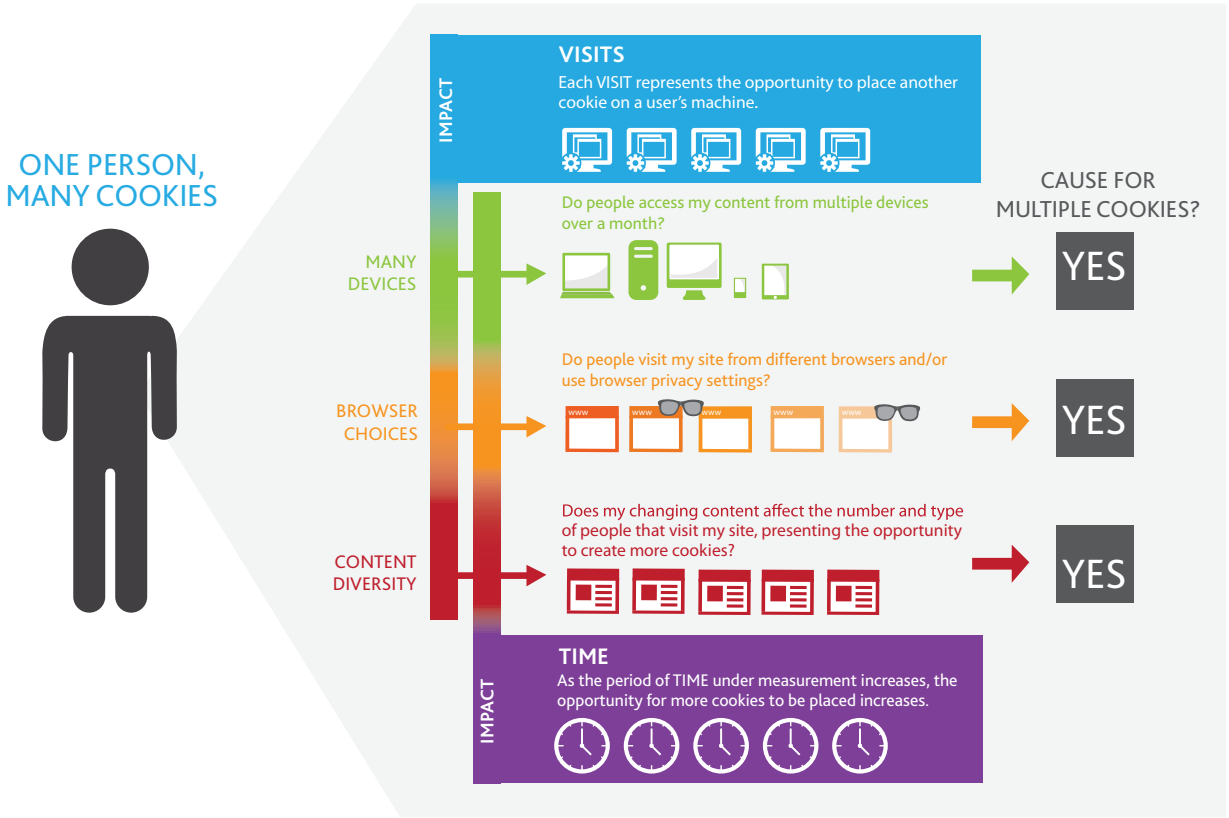
After confirming the relationship between website usage and behavior across machines, the next study examined the correlation between cookies per person on a work PC and cookies per person for demographically similar people on a home PC. The results proved less than convincing.

In tests, site-centric model assumption #3 failed when home device visitation and behavior did not predict work device observations as shown in Chart 4.

This finding calls into question the underlying principle of the site-centric hybrid methodology which holds that adjustments based on observations of individuals on one online access point can be applied to the population as a whole. Acting on this assumption carries the risk that each site may be over- or under-represented in resulting data.

### Chart 4: Visitation and Behavior on One Device (home) is Not a Good Predictor of Cookie Duplication on Another Device (work)


### Chart 5: Overcoming a Complex Relationship



Exacerbating the situation further, this study only incorporated one additional device. Because users may likely use multiple additional devices to access a site, each with unique cookie duplication rates, the complexity of adjusting for cookie duplication increases significantly, and with it, the risk of inaccuracy within the site-centric hybrid data sources.

## Confounding Effects

Individually, each assumption supporting site-centric hybrid audience calculations may be suitable for audience estimates, but the complexity of the relationship between those variables in real-world situations undoubtedly introduces interaction effects that can skew findings.

Moreover, site-centric methodologies fall short in satisfying the demands of marketers and media owners interested in quality online audience data on a number of critical criteria. First, the data does not present a consistent view of the marketplace because tagged sites are treated differently from untagged ones. Second, the data is not derived at the individual respondent level, restricting both demographic analysis and drill-down reporting capabilities. Third, relying on cookies as the independent variable represents a challenge to data accuracy and consistency on a site-to-site and month-to-month basis.

## Conclusion

User-centric hybrid online audience measurement avoids many serious pitfalls because it is not dependent on cookies. By relying equally on both panel- and census-based behavioral data, the user-centric approach offers comprehensive, representative and consistent audience information across websites. By ascribing behavior directly to measured panelists, this model bypasses cookie-related issues such as double-counting or legal restrictions on setting or reading cookies and readily accommodates future expansion as additional data sets come online.

Measurement tags will continue to play a supporting role by providing useful transaction data. This technology affords tremendous latitude for advertisers to dynamically customize online content and serve-up relevant advertising. But the liabilities associated with cookie counting offset its usefulness as a primary audience measurement tool.

By contrast, in addition to more robust, reliable data sets, the user-centric hybrid methodology was purposely built to accept more data sources, provide more detail on access points and devices, allow improvements such as additional insights from non-PC device meters and expand panel representation.

Nielsen remains committed to exploring, expanding and examining online audience measurement methodologies, in turn enabling clients to craft more efficient and effective online strategies.

## Future Outlook

Nielsen's hybrid audience measurement provides a better and more accurate way to view Internet usage on a monthly basis. As more data sources, such as mobile phone metering, become available, Nielsen will continue to enhance reporting by providing even more detail about how individuals access the Internet from those devices. The hybrid process has been intentionally designed to allow for improvements, including the potential for additional insights from metering on non-PC devices, improved panel representation and more. As always, Nielsen strives to refine its strategy for maintaining the highest quality Internet measurement data in each of the markets served.

### About Nielsen

Nielsen Holdings N.V. (NYSE: NLSN) is a global information and measurement company with leading market positions in marketing and consumer information, television and other media measurement, online intelligence, mobile measurement, trade shows and related properties. Nielsen has a presence in approximately 100 countries, with headquarters in New York, USA and Diemen, the Netherlands. For more information, visit www.nielsen.com.


i *A Guide to Understanding Online Measurement Alternatives* (A Media Rating Council Staff Point of View); MRC Staff, August 3, 2007, p9
ii Ibid.
iii *Cookie Deletion Survey Results*, Harmon (Belden Associates) , 2004
iv *Measuring Unique Visitors: Addressing the Dramatic Decline in the Accuracy of Cookie-Based Measurement*, Petersen (Jupiter Media), 2005
v *Is the Sky Falling on Cookies? Understanding the Impact of Cookie Deletion on Web Metrics*, Song (Atlas Digital Marketing), 2005
vi *The Impact of Cookie Deletion on Accuracy of Site-Server and Ad-Server Metrics: An Empirical Study*, Abraham et. al. (comScore) 2007
vii *An Analysis of Private Browsing Modes in Modern Browsers*, Aggarwal, et al. (Stanford), 2010
viii *Interactive Advertising Bureau Audience Reach Measurement Guidelines*, Version 1.0, IAB, February 23, 2009
ix Ibid.
x *Hybrid: Innovation in Panel and Census Integration*, Duterque, Stackhouse and Mazumdar (Nielsen), 2009


For more information visit www.nielsen.com



nielsen