SURVEY WARS: THE EXPLOSION OF MEASUREMENT IN AUDIENCE RATINGS

MARK BALNAVES
Professor and Senior Research Fellow in New Media
Department of Internet Studies
Curtin University of Technology
GPO Box U1987
PERTH WESTERN AUSTRALIA 6845
m.balnaves@curtin.edu.au

TOM O’REGAN
Professor of Media and Cultural Studies
School of English, Media Studies and Art History
The University of Queensland
Brisbane
QUEENSLAND AUSTRALIA 4072
t.oregan@uq.edu.au

Abstract
The ways of knowing modern audiences — broadcast ratings — is now big billion dollar business, with Google in alliance with Nielsen, one of the world’s leading media researchers, to literally auction audiences off to the highest bidder. Ironically, ‘the black box’ of survey and sampling methodology that delivers audiences remains a mystery to most, just as Art Nielsen’s original ‘black box’ — the Audimeter — for measuring radio sets baffled onlookers. The black box, though, is in crisis. It is beset by a rapidly changing media and consumer environment, heightened demand for ever more precise information, increasing measurement difficulty and declining respondent participation. Sampling has also become the subject of political and methodological battles of how to represent people, including everything from the Census to the Internet. Recent US research on non-response shows that technology rich households are rejecting participation in audience surveys. The joint Arbitron and Nielsen Project Apollo, designed to collect everything about audience behaviour also found the people were rejecting both the intrusion into their private lives and the tasks given to them. In this paper the authors outline the contemporary survey and sampling wars in modern audience ratings, including especially the proliferation of different ways of measuring audiences and audience rejection of many of them.

Introduction
Social networking has not escaped changes in audience expectations about participation in market research or ownership of their data. Facebook planned to exploit the private information of its 150 million members by creating one of the world’s largest market research databases. In an attempt to further fiscally capitalise on the explosion in popularity of the social networking site, once valued at $A24 billion, it would allow multinational companies to selectively target its members in order to research the appeal of new products. Companies would have been able to pose questions to specially selected members based on such intimate details as whether they are single or married and even whether they are gay or straight.


In February 2009, Facebook made amendments to its privacy settings that allowed them to take ownership of anything users posted onto their profile — even if they deleted their accounts.
(Walters, 2009). Users were outraged and protested against the social networking giant. Mark Zuckerberg, the founder, issued a statement saying, ‘trust us, we’re not doing this to profit from you, it’s so we are legally protected as we enable you to share content with other users and services’ (Walters, 2009). Needless to say, users weren’t happy to just ‘trust’ the CEO of a billion dollar company and Facebook reverted back to its old privacy policies. Zuckerberg subsequently set up a poll asking users to vote on new privacy policies as he still feels they need to be updated (Musil, 2009).

What we see in the Facebook experience is what history has told us about audiences and the media industry. Audiences are sensitive about how they are represented and how information about them is used. The metrics associated with audiences — how they are measured and how the information is captured — are directly related to how decisions are made about advertising, programming and the provision of services. Facebook, in short, came up against what traditional broadcasting has experienced for decades. The difference is, of course, is that individuals have a more intimate link to broadcasting of their identities through social media than they ever did in a world of scheduled programming for radio and television.

The Facebook example is important to our understanding of the active nature of modern audiences and their feelings about their position in particular kinds of communicative spaces. The authors have called this paper ‘survey wars’ precisely because there are underlying battles, or challenges, being faced audience participation in research and the quality of results that are possible if key traditional means of statistically representing audiences are eroded. Research participants in audience research, like the Facebook audiences, have expectations about how they are represented, what is going to happen with their data, and how those data are collected. Nielsen’s independent, commissioned, study on non-response bias in 2009, coordinated by Professor Peter Miller from Northwestern University, demonstrates one of the important phenomena underlying the modern survey wars — the problem of audience participation as respondents in survey panels themselves.

**Nielsen’s non-response bias study**

Nielsen conducted a non-response bias study in 2009 and its results give an insight into the role of the modern audience in survey research and the limits the audience itself puts on participation. The study consisted of 2,300 basics, with 1,000 responding households (a 95 per cent return rate) and 1,300 refusing households (a 62 per cent return rate). Special in person follow ups were made with those households that did not respond. Few studies have followed up with non responding households on why they have not participated or given data and this is what makes this particular study important. Table 1 shows the difference between responding and refusing households by day part. The p values are significant for the 8–11pm daypart, with the possibility of differences in the 5–9 am day part.

**Table 1: Daypart comparisons**

<table>
<thead>
<tr>
<th>Daypart</th>
<th>% Response</th>
<th>% Refuse</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>5–9 am</td>
<td>41.9</td>
<td>45.8</td>
<td>.11</td>
</tr>
<tr>
<td>9am–4pm</td>
<td>56.1</td>
<td>56.3</td>
<td>.94</td>
</tr>
<tr>
<td>4–8 pm</td>
<td>67.4</td>
<td>69.7</td>
<td>.46</td>
</tr>
<tr>
<td>8–11pm</td>
<td>71.2</td>
<td>76.2</td>
<td>.04</td>
</tr>
<tr>
<td>11pm–2am</td>
<td>39.3</td>
<td>39.7</td>
<td>.89</td>
</tr>
<tr>
<td>2–5 am</td>
<td>10.2</td>
<td>12.3</td>
<td>.18</td>
</tr>
<tr>
<td>12am–12am</td>
<td>91.5</td>
<td>90.1</td>
<td>.41</td>
</tr>
</tbody>
</table>

Source: Miller (2009)

Table 2 in comparison shows refusing households by Channel, with significant differences for CNN, HBO with MTV and Fox approaching significance.
Table 2: Channel comparisons

<table>
<thead>
<tr>
<th>Channel</th>
<th>% Response</th>
<th>% Refuse</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>70.3</td>
<td>73.7</td>
<td>.19</td>
</tr>
<tr>
<td>CBS</td>
<td>68.7</td>
<td>73.2</td>
<td>.12</td>
</tr>
<tr>
<td>Fox</td>
<td>68.4</td>
<td>72.7</td>
<td>.11</td>
</tr>
<tr>
<td>NBC</td>
<td>71.2</td>
<td>73.3</td>
<td>.38</td>
</tr>
<tr>
<td>Univision</td>
<td>7.6</td>
<td>6.7</td>
<td>.52</td>
</tr>
<tr>
<td>BET</td>
<td>11.3</td>
<td>12.5</td>
<td>.52</td>
</tr>
<tr>
<td>CNN</td>
<td>37.7</td>
<td>45.2</td>
<td>.00</td>
</tr>
<tr>
<td>HBO</td>
<td>17.4</td>
<td>27.9</td>
<td>.00</td>
</tr>
<tr>
<td>MTV</td>
<td>15.5</td>
<td>19.1</td>
<td>.09</td>
</tr>
</tbody>
</table>

Source: Miller (2009)

But it is the figures in Table 3 that surprised researchers and start to show why many people do not want to participate in modern audience panels and audience ratings survey panels especially. Nielsen found that technology-rich households did not want to participate in research — the very demographic that modern marketers want to grab. There are various reasons for this, ranging from an unwillingness of households to allow increased intrusion into their lives across a range of technologies and, of course, the simple fact that the technologies themselves need to be intruded upon in order to gain data (for example, additional wiring) and participants do not want their expensive technology tampered with in any way.

Table 3: Equipment comparisons

<table>
<thead>
<tr>
<th>Device</th>
<th>% Response</th>
<th>% Refuse</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Screen</td>
<td>22.6</td>
<td>41.2</td>
<td>.00</td>
</tr>
<tr>
<td>Cable</td>
<td>61.9</td>
<td>69.1</td>
<td>.00</td>
</tr>
<tr>
<td>Satellite</td>
<td>24.7</td>
<td>28.5</td>
<td>.18</td>
</tr>
<tr>
<td>DVR</td>
<td>9.7</td>
<td>20.1</td>
<td>.00</td>
</tr>
<tr>
<td>Hi Speed Web</td>
<td>72.9</td>
<td>75.3</td>
<td>.30</td>
</tr>
<tr>
<td>Web TV</td>
<td>63.6</td>
<td>69.5</td>
<td>.15</td>
</tr>
<tr>
<td>3+ TVs</td>
<td>46.1</td>
<td>60.2</td>
<td>.00</td>
</tr>
</tbody>
</table>

Source: Miller (2009)

Audience ratings panels and survey research generally have always had problems with non-response or non-participation in research, even in the days of Archibald Crossley, Hans Zeisel and Paul Lazarsfeld. However, the continued erosion of the ability to get people to participate or to give more data about themselves represents genuine difficulties in attaining quality data at the very time when there is today an explosion in demands for more measurement. Project Apollo is the paradigm example of an attempt to capture everything people do in one panel. When Nielsen and Arbitron joined forces to set up the experimental Project Apollo in 2005 they expected to capture all of the everyday behaviour of audiences, from reading papers through to using mobile phones and buying food. To their surprise they found that people did not want to participate. The more they were asked to do and to provide, the more they resisted and refused. The current situation involves many separate panels for separate activities, for example a panel for outdoor advertising, a panel for mobile phones, and so on. There is also access to ‘buyer graphics’ such as actual purchases of, say, books. The failure of Project Apollo and the recent Nielsen study, however, point to the problems of tradeoffs in the quality of data that is happening.
The fate of the panel

The survey panel was a revolution in the early 20th Century that transformed the collection of survey data. Panels are statistically representative groups of people or households that provide data to a study over time. In television ratings this longitudinal aspect has been fundamental to the operation of the business of ratings as it has provided the users of ratings the capacity to map audience flow and to get a sense of what is happening to audiences over time.

The process of selection of panels and their retention over time is therefore key to the quality of data that is retrieved. In order to discuss the mechanics of selection of panels it is worthwhile briefly revisiting the very idea of random, probability, selections. Populations in statistical sampling are operationally defined by the researcher. They must be accessible and quantifiable and related to the purpose of the research. ‘All households in Sydney’ is a definition of a population, with households as the unit of analysis. But what is a ‘household’? Is it any dwelling, including the boat in the backyard that your mother in law lives in? Is it the Hilton? If all of these examples are ‘household’, then these need to be included in the definition of the population. If a sample of households was drawn from Sydney, then ‘Sydney’ needs to be defined. Is Sydney defined by local council boundaries? Is it defined by census boundaries? Traditionally, modern audience ratings companies have conducted establishment surveys and compared them against census data. The panel for television viewing for example is then drawn from those households defined as necessary for the panel. Not all households will be chosen but those households chosen, say about 4–5,000 households in Australia, should be statistically representative of the population from which they were drawn.

When you have decided on all the definitions, then every household that you have defined as a household belongs on your list. That list is called a sampling frame. You can number each household, if you wish, and put them in a hat (a big hat in this case). Let us say that you drew out 50 households (a sample). You know that there was no bias in your choice. Each element, each household, in your sampling frame had an equal chance of being chosen. This is called a probability or random sample.

If it is not possible to get an appropriate sampling frame, or list, then the researcher can use other non-statistical means of selection or vary the definition of the population. This is often done in the case where there is unlikely to be a full accessible list of units of analysis for the study — for example there is no easily accessible list in Australia of all drug users who watch television. However, non-random, or non-probability, samples are not generalisable to the population from which they are drawn.

Television ratings panels may obtain an appropriate sampling frame but that is not the end of it. If not all the ‘basics’ — those on the list — do not want to participate, as Nielsen found in the 1990s, then the alternatives, those who say yes, may not be representative of the panel originally chosen. In the case of Nielsen’s the ‘alternates’ ended up having less televisions and were heavier television viewers than the ‘basics.’ There end up being key tradeoffs if an appropriate statistically representative group cannot be found. Table 1 provides a summary of those tradeoffs. The ultimate in survey research is gain a Data Rich and Case Rich outcome. Data Rich means that the information coming back from the audience is extensive and has qualitative depth. For example, information on Facebook is Data Rich and Case Rich outcome. Facebook has millions of users. These ‘users’ do not form into well defined and easily accessible lists. There are as a result problems in establishing or verifying identities of users and creating an acceptable list where a random draw or stratification can be made. Facebook, therefore, is Data Rich but Case Poor as an audience and as a source of information. The traditional panel for television ratings, by contrast, is Case Rich and Data Poor.
The Data Poor nature of the television ratings measured by exposure — who watches, where, for how long — was made up by the fact that users of the ratings found them practical and acceptable as a form of currency (a trade off). Miller would call this part of the ‘co-ordination rule’ in audience ratings. There was in the past confidence that the television ratings panel indeed reflected the population from which it was chosen. Through Peoplemeters the television ratings panels have provided information about an audience’s exposure to television. If a company or a researcher wants more information about whether the audience actually like programs then customised Data Rich studies are often required. Modern syndicated ratings research therefore has often complemented by customised studies seeking more detailed information about audiences.

### Table 4: Tradeoffs in the underlying survey ways

<table>
<thead>
<tr>
<th></th>
<th>Case Rich</th>
<th>Case Poor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Rich</strong></td>
<td>Single source (the Holy Grail)</td>
<td>Custom</td>
</tr>
<tr>
<td><strong>Data Poor</strong></td>
<td>Syndicated</td>
<td>Quick and dirty</td>
</tr>
</tbody>
</table>

Source: Adapted from Miller (2009)

In summary, Case Rich studies are those that use statistical sampling methods to derive their samples from sampling frames. Case Poor studies are those that have not deployed statistical sampling methods and therefore their samples cannot be used to represent the populations from which they might have been drawn. Data Rich studies are those that yield detailed data or accounts from participants. In the case of television, for example, an interviewer living with a television household for a month is getting fine-grained qualitative data. Data Poor studies are those that gain minimal information back for their purpose. In the case of television ratings, exposure gives the media industry basic information about who is watching, when and for how long. Those studies that do not gain an appropriate probability, statistical, sample and do not gain particularly rich and detailed information are both Data Poor and Case Poor. The modern problem for syndicated research is that participants do not want to … participate, especially those who are of most interest to media researchers. If panels are failing to adequately represent the modern audience, then they are Case Poor. The data, as a result is not generalisable to the defined population.

There are other factors that can affect the data coming back from the television ratings panel, for example fatigue from button pushing or simply commitments in time. These may affect the panel operation or the data but are normally accounted for.

### Alternatives to the panel (and replacing people with technology)

The broadcast audience ratings convention, or compact, has had several important components (O’Regan and Balnaves 2009). The convention at base:

1. Has exposure as the key measurement;
2. Must appeal to the inherent correctness of the measurement;
3. Uses a probability, statistical, sample;
4. Delivers a ‘single number’;
5. Is syndicated to reduce costs to subscribers;
6. Has generally been Third Party;
7. Is expected to work in the public interest (that is, accurately represent the public audience).

Media ratings systems have traditionally provided an economic foundation for advertiser-supported media. Consequently the nature of the audience measurement process affects the structure and behaviour of media companies and regulators alike. Changes in the techniques and technologies of the ratings have ‘a significant effect on the economics of media industries (because these changes can affect advertiser behaviour), the relative economic health of various segments of the media industry, and the nature of the content that media organisations provide.’ (Napoli 2003, p.65)
Although changes to the ratings convention governing audience measurement can be disruptive, these changes are driven by the inevitable gap between the measured audience and the actual audience for a service or programs. With the advent of a more diverse and fragmented media environment and fractionated audiences, increasingly demographically defined, this gap has become even more evident with the validity of ratings as currency for buying and selling media being challenged in the US. Napoli (2003) suggests that this is leading to a decline in quality and value of the ‘audience product’ — data on who is watching when — because of changes in technology and audiences.

Determining the popularity of programs remains an important reason for ratings research whether for traditional broadcasting or the Internet. Given the problems with modern panels and the use of exposure as a measure attention has turned to the ‘quality’ of the audience in addition to their size, as content providers and advertisers seek both to know their audiences/consumers better, and to hone in on those considered most ‘valuable.’ The scheduling aspect of ratings remains important in ‘traditional’ television (FTA and STV), although time-shifting whether via ‘catch-up channels’ on STV, or through use of DVRs, and the increasing availability of content online has to some extents eroded the importance of the schedule. Having said this, it is clear that scheduling decisions are still made by networks, broadcasters and narrowcasters in part at least through consideration of the ratings of a particular program, or in the case of new programs the performance of other similar programs. ‘Hammocking’ — the practice of scheduling a program before, after or between high-rating programs in order to maintain an audience — is an instance of scheduling decisions determined by ratings. This practice is however becoming more and more difficult as the audience’s availability may be dispersed across time, or determined by factors other than the push-driven schedule. Creating a market for advertising has historically been one of the principal purposes of ratings, or of high rating shows.

All the evidence points to how difficult this is becoming for television providers and media buyers as the market is more fragmented than ever. Broadcast television’s share of the advertising revenue pie is decreasing as other platforms and services attract viewers and hence become more attractive for advertisers. Neuromarketing, proprietorial registration systems, small audiences, zero ratings, fusion — all these terms or phrases signify a different type of audience ratings convention or measurement and a search for alternatives to exposure as the main metric and the panel as the mainstay for data collection. Replacing people with technology as the main reporting instrument is another option.

Aribtron the monopoly provider for radio ratings in the United States has for many years been trying to replace the paper diary with a PPM (Personal People Meter). The PPM is like a pager. It can pick up digital codes from different media and send the information to a database. Arbitron has used paper and pencil diary systems for radio since 1965. But the Australia radio broadcasting market rejects the technology push. Australia’s Commercial Radio Australia argues that the PPM:

- Costs 2 to 3 times the cost of the diary system
- Has no real evidence of long-term compliance across all demographics; e.g. once listening is lost, it is lost and cannot be regained
- Drops in breakfast listening have been recorded in trials with no valid evidence provided for device purveyors’ explanation that ‘people must have been over-reporting breakfast listening’! ‘And industry does not accept this explanation as no other daypart has been supposedly over-reported. We do not believe that tens of millions of diary keepers across the world over the last 50 years have all over-reported breakfast and no other day part.’ (Commercial Radio Australia 2008)

At the time of writing this paper, Nielsen Australia were, ironically, invited by radio broadcast networks in the United States to bring its diary into the US radio market, in competition with Arbitron. But regardless of the technology versus self report argument, if a properly constituted
panel cannot be obtained then the industry is stuck with Case Poor Data Poor. Neuromarketing and engagement are possible alternative metrics, but if there is no statistically representative sample then the industry will be relying on Data Rich Case Poor. At present, though, there no agreed upon metrics for neuromarketing and if history is a guide then it is unlikely that there can be.

Conclusion

The modern era in audience ratings is marked by what the authors call underlying survey wars. Traditional audience research has relied on statistical sampling that provided Case Rich foundations for syndicated ratings. Exposure has been the key measure and even though it is Data Poor compared with engagement it has been an acceptable tradeoff for the media industry and remains a key currency. However, the traditional panel is under threat because of problems of non-response bias, among other things. This situation is unsustainable because in the long term the media industry will end up with Case Poor Data Poor delivery of ratings. The alternatives, like engagement as a metric or drawing information from sources like Facebook, are also not sustainable. These alternatives would most likely be Case Poor and Data Rich. However, the idea of having non-representative population data for ratings would be an unacceptable tradeoff from the point of view of currency. It is one thing to make the trade off with exposure as a measure and keeping the study Case Rich and another to have data that does not represent statistically defined populations.

In the meantime, the demand for more measurement is not declining. Collecting data from proprietary set top boxes, for instance, is growing, but that faces all the difficulties of privacy and intrusion that are emerging, apart from the fact that each proprietary system will have its own measurements. Google has combined with Nielsen to try to capture the tail of small audiences in television and to auction off those audiences. Once again, though, these experiments are not delivering the types of behavioural data that panels have traditionally produced. Ironically, also, audience self report is often created to calibrate the results produced by alternative systems. This was and is the case with Arbitron and its PPM. Data is collected from people using traditional paper and pen diaries and then compared with a panel using automated systems.

This history of audience ratings for media buying has been to move away from Data Rich approaches or sources that do not deliver reliable Case Rich samples. For instance, fan mail was a major source of information about radio programs pre-Archibald Crossley’s creation of radio ratings. Fan mail is Data Rich but Case Poor. Radio ratings using exposure as a measure and reliable panels provided a stable means for collecting longitudinal data for syndication purposes. This is not to say that Data Rich sources of information are not important for audience research. However, the future of audience ratings will be dependent on solutions to the decline in Case Rich sources.

Acknowledgements

This paper is based on research from the Australian Research Council supported project The Emergence, Development and Transformation of Media Ratings Conventions and Methodologies in Australia, 1930–2008. The authors are grateful for ARC support, Professor Peter Miller’s participation in a special audience ratings seminar in Brisbane as part of the research, and reviewer comments on this paper.

REFERENCES


