Cultural Trends
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ccut20

Estimating audiences: sampling in television and radio audience research

Guy Starkey Dr

Published online: 22 Oct 2010.

To cite this article: Guy Starkey Dr (2004) Estimating audiences: sampling in television and radio audience research, Cultural Trends, 13:1, 3-25, DOI: 10.1080/0954896042000216428

To link to this article: http://dx.doi.org/10.1080/0954896042000216428

Cultural Trends
Vol. 13(1), No. 49, March 2004, pp. 3–25

Routledge
Taylor & Francis Group

Estimating Audiences: Sampling in Television and Radio Audience Research

Guy Starkey

Cultural consumption is problematic from a number of different perspectives, but certain responses from producers, regulators and commentators depend on the measurement of consumption according to quantitative and qualitative parameters. The reliability of the data can vary widely, not least because in certain areas consumption is invisible to those who would measure it, and so they must make estimates based on assumptions about methodology and sampling techniques. Whereas in auditoria, turnstiles may quite accurately quantify footfall through the premises, and sale or return inspires high levels of confidence in circulation figures for certain types of publication, broadcasters perform to intangible audiences who can be neither counted nor observed en masse. This article considers the appropriateness of sampling techniques used to produce audience research data for the broadcasting industries, for the programmers and advertisers who need detailed 'knowledge' about their audiences. It draws on the contextualization in Cultural Trends *45 (Starkey, 2003), which examined the debate around competing methodologies using either innovatory electronic devices for the measurement of consumption or more traditional human recall. The article raises important questions for those who use sampling techniques in the cultural sector and those who would interpret their data.*

Keywords: Audience research; Audiences; Radio; Sampling; Television

Introduction

Audiences in the cultural sector vary according to a number of different parameters, among them size, demographics, levels of appreciation, the nature of any interaction with the presentation and the likelihood of individuals among them returning for more on future occasions. The availability of reliable qualitative and quantitative

Correspondence to: Dr Guy Starkey, Senior Lecturer in Radio, School of Arts, Design, Media and Culture, University of Sunderland, St Peter's Campus, Sunderland SR6 0DD, UK. Email: guy.starkey@sunderland.ac.uk

ISSN 0954-8963 (print)/ISSN 1469-3690 (online)
DOI: 10.1080/0954896042000216428

data relating to consumption is key to the success of many ventures, both public and private, yet depending on the nature of the work being presented, and the context in which it is consumed, the accuracy with which audiences may be measured is another, and highly significant, variable.

When audience size has financial implications, the accuracy of measurement can be crucial to a range of different stakeholders, including investors, producers, performers, exhibitors and advertisers. Where consumption necessitates footfall, that is, audiences entering and leaving premises, individuals can be physically counted with relatively high levels of accuracy, even if the personnel or the technology required to do so on an automated basis may be too expensive for some organizations to fund from their budgets. Cinemas make returns to distributors, based on ticket sales, and the amount a feature film grosses at the box office provides data for comparative league tables, the contents of which can make and break the reputations of actors and directors. Similarly, sales of newspapers and magazines may be audited with a high level of accuracy: returns being deducted from the gross sales figures and actual sales distinguished from free copies given on a promotional basis to such outlets as hotels and airlines. Even hits on a web site can be physically counted, and information about the nature of such 'visitors' can be quickly processed and published. One caveat: neither hit nor footfall counts necessarily distinguish between repeat visitors and those who only visit once, so frequent visitors may bias the results because their characteristics outweigh the others.

Where physical enumeration is impossible, though, measurement must depend on estimation, but to be credible this has to be more reliable than mere guesswork. Far more problematic than measuring footfall, ticket sales or point-of-sale receipts is estimating the size of unseen audiences to media that are consumed in diverse locations, far removed from the various points of production and/or distribution. One example is billboards, although knowing how much traffic passes by them can provide comparative data relating to potential audiences for different sites. Because no one can be sure how many motorists and other passers-by actually do look at a billboard, the best data on consumption can only be an estimate. Audiences to other media consume their output in locations of their own choosing, which, by definition, cannot be observed. Despite the uncertainty over estimation, though, this kind of data retains an irresistible appeal to those who can benefit from it, largely because it is used to price the sale of advertising space.

For example, in the UK, the National Readership Survey (NRS) adds estimated data on consumption to the raw newspaper and magazine sales figures verified periodically by the Audit Bureau of Circulations (ABC): a single newspaper may be read by a number of people before being thrown away, and the NRS uses sampling to attempt to discover how many. ABC (who, when so commissioned, also verify attendance data at events) work with tangibles, carrying out inspections of the auditing processes by which claims of sales figures are derived, once per annum for national newspapers, larger regional newspapers and consumer magazines. NRS use personal interviews among population samples, aided by laptop computer, to produce data relating to average issue readership, reading frequency per individual, and reach over 12 months (NRS, 2003). Table 1 contrasts ABC's hard sales figures with data on consumption

Table 1 Comparison of Sales and Consumption Data for National Daily Newspapers

	Audited average daily circulation (ABC)	'Average issue' readership: individuals (NRS)	'Average issue' readership: per cent of population (NRS)	Readers per copy (column 3/ column 2, rounded to 1 decimal place)
The Sun	3,331,690	9,229,000	19.6	2.8
The Daily Mail	2,318,578	5,979,000	12.7	2.6
Daily Mirror	1,824,475	5,146,000	10.9	2.8
Daily Express	938,905	2,181,000	4.6	2.3
The Daily Telegraph	883,934	2,306,000	4.9	2.6
Daily Star	777,889	1,894,000	4.0	2.4
The Times	586,127	1,865,000	4.0	3.2
Daily Record	496,331	1,467,000	3.1	3.0
Evening Standard	373,973	1,014,000	2.2	2.7
The Guardian	333,630	1,332,000	2.8	4.0
The Independent	180,919	584,000	1.2	3.2
Financial Times	130,637	528,000	1.1	4.0
The Scotsman	72,349	220,000	0.5	3.0

Sources: ABC, 28 July–24 August 2003; NRS, July 2002–June 2003

produced by extrapolation from the responses of approximately 35,000 individuals surveyed by NRS.

However, even this further layer of sophistication is incomplete. Publishers have to rely on less sophisticated means to gather audience responses to individual items: the size of the reader response to competitions, problem pages and telephone polls, for example, but they do not know which articles have been read and which were not, and their advertisers may never know if their messages have been assimilated or ignored, unless they add identifiers to their campaigns which encourage specific responses, such as to special offers.

The data from the ABC audits may reasonably be read as broadly confirming the estimates behind the NRS figures (even though the periods surveyed are not an exact match). None of the NRS figures for readership of national dailies lie outside a tolerance of common sense assumptions about the number of people who are likely to have access to a typical copy: too many readers per copy would suggest widespread sharing of newspapers at workplaces and elsewhere, to an extent that may stretch credibility. A much lower 'readers per copy' figure would suggest significant newspaper sharing within households does not happen. Interestingly, the newspapers with the smallest circulations would appear to be shared more than those higher up the table. This could be an accurate reflection of patterns in newspaper sharing, or a developing statistical error as the numbers of actual readers per publication discovered among the NRS sample declines. It is pertinent to consider why on average nearly twice as many people should read a single copy of the *Financial Times* as would read a copy of the *Daily Express*, particularly when NRS also reports that the ABC1 demographic group outnumbers C2DEs among the *FT*'s readership by a ratio of 14:1.

It may be significant that above 750,000 circulation, newspapers cluster around 2.5 readers per copy, whereas the rest range from 2.7 to 4.0 readers per copy (with standard deviations before rounding of 0.19 and 0.51 respectively).

Broadcasters wanting audience data, however, must rely entirely on estimates produced by sampling. Unlike in the print industry, there is no hard copy audit that may be carried out to confirm viewing and listening, because their audiences are invisible to them: disparate individuals consuming their products in their own homes, in cars, on public transport and elsewhere. The ubiquity of the broadcast media renders observation of their whole audiences impossible (Tryfos, 1996, p. 320), and their situation unique in the cultural sector in terms of sole measurement methodology. *Cultural Trends* 45 (Starkey, 2003) examined the use of both respondent recall and electronic metering among sample groups of listeners and viewers to produce audience data for radio and television, and observed that different measurement methodologies tend to produce different results (Starkey, 2003, pp. 66–67). This article explores the use of sampling for estimation through extrapolation, and asks whether the faith placed by broadcasters, regulators, advertisers and commentators in the data it produces is misplaced.

The article is written in six parts, which respectively consider the principles of sampling; sampling in practice; using and misusing estimates from sample data; sampling in crisis; selecting samples; and the conclusions that can be drawn about estimating audiences in television and radio research.

Part 1: Principles of Sampling

For the production of regular quantitative audience data, both the Broadcasters' Audience Research Board (BARB) and Radio Joint Audience Research (RAJAR) use non-random sampling methods, positively selecting 'types' of respondents within different populations defined by region or Total Survey Area (TSA) respectively. Random sampling, in which respondents are selected completely by chance from the various populations, would take account only accidentally of demographic trends within each population and be more likely to invite criticisms of being unrepresentative, although Tryfos's brief history of sampling indicates this to have been the most popular method of sampling for most purposes from as early 1925 (1996, p. 14). Much of the literature on sampling methodology continues to advocate completely random sampling, in which purely mechanistic means are deployed in order to generate a sample set, contending that this randomness should ensure the entire population is accurately represented in the sample, anyway, without the need for researcher intervention (Stopher & Meyburg, 1979, pp. 14–19).

Truly random selection of respondents is itself problematic, because to be efficient the process must ensure that each unit has an equal chance of being selected (Stopher & Meyburg, 1979, p. 24), for only then will selection bias be completely eliminated. Choosing respondents randomly from telephone directories, for example, would exclude all those households without telephones, or which are ex-directory. Electoral roles are notoriously incomplete and often already dated as soon as they are published,

because people move house and forget to inform the electoral registration officer. Commercial databases compiled for marketing purposes often relate to responses to other marketing initiatives and are therefore by definition incomplete.

Done appropriately, though, random selection is widely seen as the only way to completely avoid selection bias, in which characteristics might be introduced into the sample, which do not adequately reflect the population (Stuart, 1962, pp. 10–13). Tryfos described a 'reexamination and a reevaluation of the foundations of random sampling', though, beginning in the 1980s (1996, p. 14), inspiring new interest in selection as a means of sample construction. Stratified random sampling, as a compromise between the two positions, allows samples to benefit from randomness in the choice of individual respondents within particular groups or strata (defined by geographical, demographic or other criteria) but ensures a degree of representativeness in terms of the numbers of individuals occurring within each of the strata. This produces actual, as well as common-sense gains in the precision of sample design (Stuart, 1962, p. 44) in ways that are appropriate to social scientific research, because people vary according to a range of socio-economic criteria far more than do, for instance, individual plants in a crop survey, where the research focus may be purely on incidence, rather than behaviour.

Counting plants in randomly selected plots within a field of several acres is more likely to produce accurate estimates for the whole field if the soil conditions vary only slightly. Over larger areas, even for crop surveys, as greater differences in terrain occur, stratification becomes a useful tool in weighting samples according to identifiable variables (Som, 1973, p. 138). Human settlement tends to be in clusters, where ethnic minorities, for example, might be concentrated in some places but not others. Differentials in the cost of housing also determine the distribution of people according to wealth, professional status, class and age, making weighting through stratification even more appropriate for this kind of survey, and Tryfos (1996, pp. 96–99) demonstrates mathematically why a proportional stratified sample is 'nearly always better' than a same-size simple random sample. The caveat lies, of course, in the nature of the stratification (Yates, 1981, pp. 26, 264): if the strata are inappropriately defined, or the weighting given to them incorrect, more accurate results may well have been obtained by a simple random survey, providing the sample was large enough.

Sample size is governed by that of the population and the nature of stratification as it impacts on the structure of the sample. Particularly because cost is a factor in any business, researchers usually resort to conventional sample sizes for common tasks. A national opinion poll of the whole UK, for instance, will usually use the commonly-accepted sample size of just over 1,000 respondents, but polls of between 1,000 and 1,500 are standard in the USA for a population of more than four times as large (Thompson, 1997, p. 65). Such small samples are considered acceptable when sufficient precision can be achieved for the estimators required: determining the overall distribution of voting intention in a national political context offering two main choices, Democrat and Republican, can be achieved on such a small sample, whereas producing reliable data on the geographical and demographic distribution of those votes becomes more problematic, given its size. Generally, though,

anyone considering audience research data should heed the warning by Tryfos that estimation by sampling is not a form of magic, which can be considered infallible: 'there are no known methods . . . which ensure with certainty that the sample estimates will be equal to the unknown population characteristics' (1996, p. 15).

The stratification needed to produce the kind of data sought for audience research is highly complex: populations being defined by not one demographic descriptor, but by several. There may be other considerations, such as for BARB the six different television viewing platforms: namely analogue and digital versions of terrestrial, satellite and cable respectively. In the UK analogue terrestrial viewing still predominates, but consumption of cable and satellite channels is growing, albeit fragmented across the different alternative platforms. In 2003 RAJAR introduced data for DAB (Digital Audio Broadcasting) into its survey (2003a). Common sense dictates that robust viewing data for the London region, for example, would require there to be many more than one professional 18–35 year-old male from a significant ethnic minority using Freeview to watch television, but how many should there be? Stopher & Meyburg (1979, p. 33) are not alone in advocating sampling 5 per cent of the number of units (in this case, individuals) in each stratum. Increasing the number of strata or the number of layers of stratification, then, may necessitate an increase in the size of the sample, although it may already be large enough to accommodate such changes.

As both BARB and RAJAR publish data for national audiences in addition to those in specific territories within the UK, they attempt to construct samples that reflect the demographic structure of each of the respective populations they are supposed to represent, both national and regional or local. BARB produces separate data for 14 ITV regions and 14 BBC regions, while RAJAR services over 270 different TSAs, making its *Quarterly Summary* particularly complex (O'Hara, 2003, p. 81).

BARB's methodology involves placing electronic measurement equipment with a panel (that is, remaining broadly constant, rather than changing every week) of 5,100 households, sub-divisible by region. The television industry wants, and is prepared to pay the extra costs of, the 'overnight' data about yesterday's viewing, which

Table 2 Sample Sizes Used by RAJAR to Survey Different-sized Populations

Population aged 15 + (000's)	Commercial Reporting sample	Based on	BBC Radio Reporting sample	Based on
10,000	2400	Quarter	1000	Quarter
8,000–9,999	1900	Quarter	1000	Quarter
6,000–7,999	1400	Quarter	1000	Quarter
4,000–5,999	1000	Quarter	1000	Quarter
1,750–3,999	1000	6 months	650	Quarter
1,000–1,749	1000	6 months	1000	6 months
300–999	650	6 months	650	6 months
Below 300	500	12 months	650	12 months
Opt outs	300	12 months	300	12 months

Source: RAJAR, 2003b

is made possible by the technology used. By contrast, within a national total of 226,185, RAJAR recruits samples of between 300 and 2,400 individuals, depending on the size of the particular population, to each complete a listening diary over a week. Table 2 shows how cost limits the preparedness of RAJAR and its subscribers to recruit respondents in smaller TSAs. Selecting, recruiting, briefing and debriefing each respondent adds to the annual £4 million cost of producing the survey, and for smaller stations to participate in it, their individual contributions have to be affordable. This, in turn, constrains the production of data in smaller populations, the data 'rolling' across quarters in order to even out the inconsistencies that would be likely to result from changing small samples wholesale on each occasion.

Part 2: Sampling in Practice

Sample error is considered inherently unavoidable in the use of relatively small numbers of cases to determine the nature of whole populations, although it is considered to be within acceptable tolerances when random, as opposed to constant (Deacon *et al.*, 1999, p. 42). Random errors in this instance would be characterized by the effect of maverick listening or viewing by a single respondent choosing to consume a programme that would be an atypical choice of most people in the demographic group the respondent has been chosen to represent. One effect of random error could be the chance promotion or demotion of a programme, channel or station in a league table, with potentially beneficial or catastrophic results, respectively. This would be very hard to eliminate without micro-management of the resultant data, and editorial decisions would be needed, over the extent of maverick activity and how to compensate for it, with the risk of overcompensation where what is perceived as atypical activity may actually be more typical than assumed by the research organization.

Constant errors, though, systematically distort data because they do not occur randomly, but may derive from inappropriate selection methods, repeatedly overemphasizing one or more elements of the population (Som, 1973, pp. 278–280) and hence, skewing results, as might be the case with the NRS figures producing surprising 'readers per copy' data for smaller circulation newspapers in Table 1. Errors may also result from systemic faults in survey processes and practices, or even at the later stages of interpretation and publication. Sometimes broadcasters receiving infelicitous survey results claim inappropriate sample design to be disadvantaging them during data collection. Just as with niche publications, targeting small specialist audiences, some radio stations and minority cable and satellite television channels can feel that even weighted samples risk misrepresenting their audiences. They say that either the weighting is inappropriate or that part of the sample intended to represent their target demographic is so small that the probability of their consumption choices being unrepresentative is unacceptably high.

Avtar Lit, Chairman of the Asian station in London, called Sunrise Radio, considered that stations such as his had been underrepresented in RAJAR's sample for years (War!, 2002). Paradoxically, though, the experience in Leicester, of the commercial

station Sabras Sound and the BBC Asian Network, both of them broadcasting on the unfashionable AM waveband, has been much happier: in the third quarter of 2002 they were the 7th and 10th most listened to stations in the Leicester Sound TSA respectively, Sabras being ranked ahead of Classic FM, Radio Five Live, TalkSPORT, Virgin and Classic Gold Gem (Boon, 2003, p. 22). Lit's however, has not been a lone voice: Nigel Reeve, as Head of Fusion Radio Group, called for reach data to be calculated as a percentage of a station's target audience, rather than out of 'all adults' (2002). His analysis noted that the first Independent Local Radio (ILR) stations in the UK were deliberately of wide appeal, most offering a 'full service' that included minority programming in off-peak periods in the absence of competing full-time stations targeting those particular demographics or interest groups.

When Classic FM launched in 1992, it was specifically targeted at an ABC1 audience, programming only 'classical' music. Today's crowded radio market includes many stations who with their whole output target specific groups according to ethnic, age or musical taste criteria, and figures demonstrating their ability to reach those communities may be more appropriate and of greater interest to advertisers. Reeve cited the case of Fusion 107.9 FM in Oxford, targeting a youth market with 'cutting edge' music, and achieving a 10 per cent reach among the target market but only 3 per cent of all adults. Their 'short-term' goal of 20 per cent in the target market would translate into only 6 per cent of all adults, and Reeve observed that such a reach figure, quoted widely, would unfortunately create an impression of failure, rather than success.

Instead of just asking for a change in reporting, though, Reeve argued for changes to the weighting of RAJAR's samples in order to place more listening diaries with the demographic groups targeted by the various niche stations in the commercial radio sector, in order to 'show the success levels' of the industry. Practically, this could be done, but only to produce a second set of results within smaller populations, those ethnic, age or interest sub-groups identified by Reeve. Concentrating diaries in this way, and including the data from them in the top-line figures for larger populations would constitute selection bias and so significantly skew all-adults results in favour of the minority stations, because of the constant error they would introduce into those samples. In fact, as is often the case with complaints about audience research methodology, Fusion 107.9 FM's low reach figures were probably also symptomatic of either unappealing programming or overwhelming competition. In the RAJAR survey for the period, the second quarter of 2002, the station's average listening hours figure per week per listener (as opposed to all adults) was reported as only 1.8. This was a particularly poor performance, compared with Oxford's rival mid-market station, FOX FM, which achieved weekly listening hours in its TSA of 10.3 per listener on a reach of 36 per cent (RAJAR, 2002).

At fault must have been either Fusion's performance (and marketing may be as much to blame as the programming) or the sample design used in the survey, or, of course, both. To believe it was the sample, is to contend that a significant number of diaries were inappropriately placed, with either non-listeners or light listeners to

Fusion. Whichever it was, the station subsequently re-branded itself as Passion 107.9 FM in early 2003. There are clear synergies between Avtar Lit's position on Asian respondents, as well as that of Nigel Reeve on youth, and the importance to the Welsh-language services, S4C and Radio Cymru among them, of having an appropriate number of Welsh speakers in appropriate strata of the samples used in their coverage areas.

Certain demographic groups may, by their lifestyles, preclude themselves from accurate representation in samples. Large numbers of young people, among them students, live without parents in very temporary accommodation, which, like that of some ethnic minorities, including asylum seekers, is likely to fall outside researchers' definitions of 'households'. In the UK, for instance, BARB and RAJAR only survey 'private households', yet there is no discernible demand from other stakeholders for more inclusive data. This probably is due to cost factors, but it could be read as if the broadcasting industry and advertisers are content to collude in the disenfranchising of certain groups from influencing broadcasting by their not being represented in the production of audience data.

Researchers seeking respondents according to quotas (in order to complete strata), as opposed to calling in response to randomly generated lists of addresses, may avoid certain neighbourhoods because they would feel threatened or insecure in them, while some demographic groups may on average be more disposed than others to attempt accurate recording of listening in written diaries or faithfully logging in and out to register their presence in front of a metered television set in response to the inducements offered them in return for their participation. It may also be that mere willingness to comply with research organizations' requests to join a sample may be a characteristic of people with greater tendencies towards the consumption of certain types of programming, rather than others, although survey organizations such as GfK do report very high response rates, as much as 90 per cent, among those they approach (Electronic measurement, 2003). Perhaps BARB and RAJAR are largely surveying 'survey nerds': if they do not start out as such, perhaps they become them, particularly in the case of long-term membership of a panel. Conversely, depending entirely on new recruits each week may mean RAJAR only measures listening by atypical 'survey virgins', suddenly aware of the range of choices on offer, tempted to experiment and who have yet to settle back into more representative patterns of consumption.

The ability to report accurately, where recall of past events or making entries in a listening diary is concerned, may also be a characteristic of some demographic groups rather than others: Frankel (1969) reported that a telephone survey of contiguous listening to radio stations in New York produced responses that were only 91 per cent accurate. That is, when prompted by the interviewer to identify which stations they were currently listening to, 9 per cent of respondents failed to do so correctly, when compared with the interviewers' compliance monitoring of the output of the 20 most popular stations in the city. There is little compliance monitoring of the respondents in most surveys: although quality checks are carried out on, for example, nil returns, no one routinely accompanies RAJAR's diary-keepers to ensure the information they record is correct, just as no one watches BARB's metered minority

of television viewers to ensure they are watching when the meters say they are. Any of these issues can result in constant error, present and actively skewing results, but undetectable without further research.

Because it is based on mathematical assumptions of probability, the logic of sampling methodology argues that random errors, in this case maverick and so, 'unrepresentative' consumption, will reduce in their effect on overall results as the sample size increases: a single maverick in a thousand respondents will cause a greater error than one in 2,500. Statistical estimations of inaccuracy decrease as the sample increases. However, and perhaps surprisingly, a point can be reached where further increases to sample size produce only very small improvements in accuracy (Henry, 1990, p. 118). Notwithstanding the logistical demands of using larger samples, this widely accepted principle enables the audience research organizations to limit the cost of fieldwork, while claiming a high degree of accuracy. Mathematical models for calculating the optimum sample size for balancing cost against accuracy consider the desirability of obtaining supplementary information: asking at what point is the benefit of getting that information negated by the additional cost required to do it (Stopher & Meyburg, 1979, pp. 93–98).

Much as this may displease those who are disadvantaged by its results, there is no practical way to definitively 'disprove' the accuracy of audience research that has arguably been appropriately carried out by sampling, although doubts will inevitably be greater over results produced with smaller samples or findings about minority activities, such as those in Table 1 with regard to the press. It would obviously be prohibitively expensive to measure actual viewing or listening by surveying every member of the UK population and comparing these new results with the routine extrapolations by, for instance, BARB from 5,100 metered households to otherwise invisible television audiences of over 24 million households. Because the audience data produced by both BARB and RAJAR normally enjoy wide support among the major stakeholders, the orthodoxies of audience sampling are rarely challenged, and the media industries are accustomed to working with, rather than against, the data. Usually they have no choice, because of the importance placed on the data by others. For example, in October 2003 the release of RAJAR's headline third quarter data caused the share price of 95.8 Capital FM's parent company to fall by 7 per cent, after the London station 'suffered its worst ever loss of listeners' (Goodway, 2003).

Table 3 shows a common use of BARB data in 'league tables' of television programmes, in this case in the broadcasting industry press. Comparative tables can make fascinating reading among the professionals who read the trade press, as well for the general public: one soap opera falling behind another in the 'ratings war' being likely to generate sensational headlines in the tabloids. Tryfos considered broadcast audience data to generate more interest than any other (1996, p. 320). The viewing data for individual programmes contribute to the relevant channels' overall share figures for the week. The strong performances of ITV1's staple soaps *Coronation Street* and *Emmerdale* alone secured the channel 11 of the places in the top 20, but with the bottom half of the top 60 dominated by programmes on BBC1, in this case the overall weekly share figures were, in percentages: BBC1 25.7, ITV1 24.8, BBC2 11.5, Channel Four 9.1 and five

Table 3 BARB Data for the Top 20 Programmes on Terrestrial Television Channels, Ranked by Audience Size, Week Ending 14 September 2003

This week	Last week	Title	Viewers (millions)	Share (per cent)	Channel
1	1	EastEnders	13.65	54.34	BBC1
2	3	Coronation Street	13.62	58.63	ITV1
3	2	EastEnders	13.01	57.24	BBC1
4	5	Coronation Street	12.68	48.80	ITV1
5	8	Coronation Street	12.51	56.47	ITV1
6	9	Coronation Street	12.00	57.54	ITV1
7	4	EastEnders	11.97	56.20	BBC1
8	7	Coronation Street	11.85	51.44	ITV1
9	6	EastEnders	11.76	53.47	BBC1
10	18	A Touch of Frost	10.37	43.00	ITV1
11	11	Emmerdale	9.90	48.68	ITV1
12	13	Emmerdale	9.70	48.35	ITV1
13	14	Emmerdale	9.65	47.02	ITV1
14	12	Casualty	9.17	40.35	BBC1
15	15	Casualty	9.05	36.17	BBC1
16	16	Emmerdale	8.91	45.83	ITV1
17	19	Emmerdale	8.86	48.74	ITV1
18	21	Suspicion	8.56	37.70	ITV1
=19	24	Emmerdale	8.49	42.95	ITV1
=19	17	Holby City	8.49	34.97	BBC1

Source: Ratings, week ending 14 September. *Broadcast*, 2003

6.2. So, because they are aggregated with audience data for other programmes, the viewing figures for even the lower-ranking shows wield an importance that escapes the attention of the headline writers. Yet, as was suggested by Table 1, it is likely that data derived from samples reporting viewing programmes in much smaller numbers are less reliable than those hits that contribute to the higher positions in the league table. In other words, the random error that is more probable to occur over estimating audiences for minority programming may not be sufficiently self-correcting to prevent distortion to more general data, such as channel reach.

For example, in that same week, the BARB estimate of the audience for BBC1's Saturday evening news bulletin was 4.58 million viewers, placing it at 50 in the league table. Table 4 shows how in a typical week the number of individual respondents in BARB's sample of 5,100 UK private households equipped with television sets is 11,614, and that in a national population of 55.821 million individuals, the ratio of actual individuals to survey respondents would be 4,806:1. That is, each respondent would represent 4,806 actual viewers, a figure which is very similar to the ratio of actual television households to metered households in the sample, namely 4,877:1. It would appear, therefore, that the programme's estimated audience size, share and rank order were entirely due to electronic reports from 953 respondents who were logged in at the time, in 425 households. Had by chance BARB placed the meters producing those results in 425 different homes meeting all the same

Table 4 Ratio of Sample Size to Population, National Audience Surveys Reporting in Mid-2003

	BARB	RAJAR	GfK	NRS
Population size	55,821,000 24,873,000[a]	49,029,000	45,036,000	47,407,000
Sample size	11,614 5,100[a]	226,185	2,080	34,931
Ratio population:sample	4,806:1 4,877:1[a]	217:1	21,652:1	1,357:1
Stated population definition	UK aged 4+ individuals/ [a]households	UK adults aged 15+	GB adults aged 16+	GB adults aged 15+

Sources: BARB, 2003a; RAJAR, 2003f; GfK, 2003; NRS, 2003

demographic criteria, how confident can we be that the same total number of individuals in them would all have tuned in to BBC1 at the time of the news bulletin? A very robust confidence in sampling techniques would be needed to assert that non-random samples are so reliable that every single one would behave in exactly the same way as BARB's actual respondents. The far greater probability is that, if the meters were in 425 different homes, the programme would appear lower down the table, unless, of course, more individual respondents watched in those households which tuned in, than in the original 425.

It is not just broadcasters who find this kind of data irresistible. Advertisers need to target their advertising at television channels and radio stations that attract the audiences they want to reach, otherwise the money they pay the broadcasters for the airtime will be relatively ineffectively spent. Broadcasting is also time-sensitive, in that certain programmes or particular dayparts will attract different audiences in demographic terms, and advertising a new hairdressing salon on a Saturday afternoon may be particularly wasteful if a predominance of football commentary and results attracts disproportionately small numbers of women to a particular radio station (Tryfos, 1996, p. 321).

Populations being surveyed for the purposes of audience research are rarely static. For a sample to remain representative, it must be adjusted in line with demographic and other changes occurring in the population. Beyond the more obvious movements of people into or out of a TSA or a region, there are effects of both migration and ageing over the whole of the UK to be accounted for. Where delivery platforms are an issue, and in both television and radio they will be for the foreseeable future, the sample must also represent digital viewers and listeners of each type in appropriate numbers. Table 4 shows how the assumptions made by different survey organizations differ in quality, in terms of the size of the population represented as the potential national audience and the sample required to produce estimates within the tolerances discussed above.

The fourth column relates to a new national monthly survey of television and radio audiences by GfK, begun in May 2003 and using a controversial electronic

'Audiometer' worn by respondents to measure exposure to different services (Starkey, 2003, p. 65). Differences in sample size may be entirely attributable to the complexity of the surveys concerned, and will almost certainly impact upon the cost of managing them and the data they produce. GfK's survey restricts reporting to national audiences and the Greater London area. RAJAR's sample appears by far the largest, although this is partly due to rolling across survey periods (see Table 2), and NRS uses the second largest because it, too, has a large number of products to survey: being approximately 260. Both the RAJAR and NRS surveys compare favourably with the sample of 21,000 respondents used in the *UK 2000 Time Use Survey* by the Office for National Statistics (2000) to produce data for a number of government agencies and the Economic and Social Research Council, as discussed in Allin (2003, pp. 85–91). RAJAR's sample is more than 10 times the size of the ONS survey and NRS's nearly twice as big, so despite the accepted wisdom of sampling principles discussed above, the data produced through estimation for national print and radio products enjoy a robustness that is not shared by either BARB or GfK, whose population-to-sample ratios are far higher, despite GfK's assertion that 'robust sampling design ensures that the sample is nationally representative of the population of Great Britain' (GfK, 2003).

Such confidence does not extend, either, to RAJAR's smaller TSAs or the more specialized publications being surveyed by NRS, such as *Angler's Mail* and *Pregnancy and Birth* over which each incidence of random error can have a greater influence. As BARB Chief Executive, Caroline McDevitt put it, 'the smaller the ratings, the greater the sampling error' (2002). Doubts over the GfK survey stem not from its sample size, but primarily from the electronic metering system it uses to gather data from respondents: in October 2003 RAJAR announced a second series of tests on meters (2003c), but even if the new technology were found to produce identical results to the traditional recall diary with the same respondents, expecting to derive even broadly similar results by using different methodologies with different population samples would be naïve, as the BARB experience described below would seem to confirm. Table 5 shows how strikingly GfK reach data produced in the third quarter of 2003 differs from reach data produced by both RAJAR and BARB. There are of

Table 5 Comparison of Television and Radio Reach Data from BARB, RAJAR and Gfk (Percentages), over the Quarter 23 June–14 September 2003

Position	BARB	GfK (for television)	RAJAR	GfK (for radio)
1	BBC1: 86	BBC1: 96	Radio 2: 26	Radio 4: 40
2	ITV1: 82	BBC2: 88	Radio 4: 20	Radio 2: 36
3	BBC2: 74	ITV1: 88	Radio 1: 20	Radio 1: 29
4	C4/S4C: 72	Channel 4: 81	Classic FM: 13	Five Live: 20
5	five: 51	five: 64	Five Live: 12	talkSPORT: 14
6	others: 45	Sky Sports 1: 21	Virgin (AM/FM): 6	Classic FM: 13
7		Sky News: 20	Radio 3: 5	World Service: 11
8		Sky Sports 2: 19	talkSPORT: 4	Radio 3: 9
9				Virgin (AM/FM): 9

Sources: BARB, 2003b; RAJAR, 2003d; GfK, 2003

course methodological and some minor reporting differences between the surveys, for example RAJAR define reach as listening for five consecutive minutes, and BARB define it as three, rendering comparison problematic (O'Hara, 2003, p. 82). The Audiometer used by GfK has to produce a positive match for two 4-second bursts of radio or television sound against a control recording, to contribute to the reach data for an individual station or channel. However, the main point in comparing the data is simply to show how the headline figures in different survey organizations can vary from other accounts, although they each claim to present an 'accurate' picture of viewing and/or listening.

The use of volunteers in samples where they are expected to act positively in order to generate data, (by, for example, filling in a recall diary, or registering their presence in a room with a television by keying a code into a keypad) carries a high risk of constant error, which cannot be detected or adjusted for. Collett and Lamb's ethnographic study of households with set meters (1986), (in which they placed cameras inside television cabinets to record what the viewers were doing while being counted as watching the programmes on the screen) identified a range of domestic and socially interactional activities being performed by respondents, which reduced the consumption of television output to a secondary status—just as radio is often perceived as being consumed merely as a background to other, primary tasks. More recent research by the London Business School (Ritson, 2002), albeit in only eight households, questioned recorded 'viewing' of commercial breaks because the subjects observed paid the least attention to the advertisements, even when they were still in front of the television.

In addition, Beckett (2001) observed systematic misbehaviour in one BARB household, where they reported having been told by the equipment installation engineer that 'no one' bothers to log out when leaving the room temporarily, where they tended to turn the television volume down and listen to the radio while the meter was logging viewing, and where an unmetered set in the kitchen was used for watching adult programmes in the late evening on Channel Five. He also found evidence of respondent reflexivity (individuals distorting the data they themselves produce) in the way individuals may contribute to the survey. Despite assurances from BARB that anyone working in the industry would be eliminated from their survey, he knew two television journalists who were panel members. The self-conscious confession of unmetered viewing of 'adult' material in one household accords with Sabo's observations (2002) about listeners reporting listening to 'cool' radio formats while concealing listening to less politically correct material, as previously discussed in Starkey (2003, p. 61).

Given the potential effect on both the BARB and RAJAR surveys, of individual instances of maverick viewing and listening, it is hard to imagine a stakeholder in one broadcaster or programme never choosing to make atypical choices in order to influence the results of the survey for which he or she is providing data. Others who are disinterested but also perceive the importance of their role as respondents may choose to inflate or deflate results for one programme or service, for reasons as diverse as friendship or kinship with its employees, approval of particular presenters

or genres, or even protest actions against controversial programming, just as a self-appointed moral crusader might have boycotted all of the first incarnation of Channel Five, in response to the inclusion of late-night soft porn in its schedules. The very notion of individuals 'representing' thousands of others in contexts such as these renders atypical behaviour highly significant, yet almost totally undetectable.

Part 3: Using and Misusing Estimates from Sample Data

BARB's estimated data inform the process by which television producers and commissioners alike plan their schedules and the production required to fill them, as much at the BBC as in the commercial sector. Outside peak time, many of those decisions will inevitably be based on the false assumptions that derive from using less robust data. In radio, too, it is a common practice to extrapolate conclusions from RAJAR's figures that lack the certainty that can derive from considering larger audiences to more popular programmes. BARB, RAJAR and NRS each produce more detailed audience survey data, which generally receive less publicity than the 'top line' figures that make newspaper headlines and form the basis of circulation claims by rival producers. Often, the data are available only to subscribers, and in the case of BARB and RAJAR, it takes the form of audience analyses by time segment. As with the BBC1 news bulletin discussed above, outside peak times, and, inevitably, in the smallest of radio TSAs and television regions, where the numbers of respondents listening or viewing is comparatively very low, conclusions about their demographic nature can be very tenuous indeed. As Stopher and Meyburg put it (1979, p. 10), 'the sample size and the conduct of the sampling process have profound effects upon any subsequent multivariate analysis of the data collected'.

For example, even a regional commercial radio station with a sizeable TSA of between 4 million and 6 million is surveyed via a sample of only 1,000 respondents in a quarterly reporting period (Table 2). A weekly reach figure for the station as a whole of 25 per cent would mean only 250 respondents in the local sample reported listening to it at all over a seven-day period. Of those 250 listeners, much smaller numbers will have been tuned in at times other than the breakfast peak period, mid-morning or afternoon drive time, so using RAJAR's data about their demographic nature in sales pitches to advertisers is stretching the validity of the survey to its utmost. Advertisers receiving such approaches should beware the easy conclusions the data seem to suggest.

An example is Galaxy 105, based in Leeds and broadcasting to the Yorkshire region. In its TSA of 4.192 million it achieved a weekly reach of 25 per cent in the second quarter of 2003, being an estimate of 1.034 million 'listeners' (RAJAR, 2003e). So, each listening respondent is considered to represent 4,136 actual listeners, a much less convincing ratio than the 217:1 for the national stations surveyed (Table 4). Although this station avoids making extravagant claims for the research on its web site, it does point out that in spite of its mix of popular 'dance and r "n" b' music being aimed at 15–34 year olds, 61 per cent of its listeners are 25 or over (Galaxy, 2003). In fact, all the station can really know from the RAJAR estimates is

that 153 members of the sample are over 25 years old (that is, 61 per cent of the 250 reporting listening to Galaxy in a week). That 153 people surveyed in Yorkshire should listen to Galaxy for at least five minutes in a given week and happen to be over 25 is much more likely to be a chance result than one that is infinitely repeatable with different non-randomly selected samples meeting the same demographic criteria in the station's TSA.

A common misuse of such data is to draw sweeping conclusions about listening choices, based on which subcategories of listening respondents are best represented at certain times and on certain days. That a particular programme or daypart attracts a higher proportion of ABC1 listeners, for example, may be an indication of preferences in the wider population, but because the sample is being sliced into ever-smaller pieces, the reliability of the data the process produces decreases considerably. However, because of the invisibility of consumption of the medium, radio programmers have little else, other than anecdotal evidence, upon which to base their decisions about content and strategy. Therefore, experienced professionals will use the fine RAJAR data as a single indicator, to perhaps confirm or dispute subjective evidence from colleagues and listeners met informally, the postbag, and responses to phone-in competitions or invitations to text the studio, rather than relying entirely on survey data. Frequently, of course, it will confirm assumptions, such as male dominance of the Saturday afternoon sports programme, because of the concentration on football.

Sometimes, particularly when marketing stations, the imagination takes over where the research finishes: for instance, Emap described the listeners to its Big City network of eight radio stations in the north of England as being: 'spirited, independent, sociable and confident and metropolitan. They lead an active social life reflecting their main agenda which is to have a good time' (Emap, 2003).

Part 4: Sampling in Crisis

The very public controversy that resulted in a temporary collapse in confidence in the BARB system in January 2002 (Wells, 2002), arose from the first complete change of the panel of respondents in over 30 years: a move intended to 'future proof' the survey in response to audience fragmentation due to developing platforms, new technology (such as Personal Video Recorders) and hence changing viewing habits. Announcing to the press that the new panel was to be 'more geographically and demographically representative', BARB had intended that more households would be surveyed in London and that the sample would include younger viewers, but the industry seemed unprepared for the sudden changes to the results, that such a move was bound to produce. Parallel tests in the preceding December had already shown that the new panel was watching television around 5 per cent less than the old one (Deans, 2002a), and of course, their viewing habits also differed from the old.

So great was the difference, that BARB suspended publication of the results for two weeks, explaining that the parallel testing of the new panel had begun late, on 20 December, and that it had not yet had the four weeks it had anticipated would be an appropriate 'settling down' period (Gibson, 2002). Broadcasters and advertisers

alike struggled without their daily diet of overnight viewing figures, and the affirmation or otherwise of programming and airtime buying decisions they provide (Deans, 2002a). When publication resumed, BARB were only using data from 3,800 households, rather than the entire, newly enlarged panel of over 5,000 they had anticipated (TV ratings are back, 2002). Among the headline figures were a number of surprises:

- While total viewing fell marginally, viewing of the commercial channels fell by 13.5 per cent compared with January 2001.
- The biggest shock in the new figures was for ITV1, the adult audience for which dropped by 25 per cent.
- Viewers in the higher-spending ABC1 demographic group were watching commercial channels 15 per cent less and ITV1 26 per cent less.
- Channel 4's key target audience of 16–34-year olds fell by 38 per cent, compared with an overall drop across all commercial channels of 21.5 per cent.
- Channel 5's ABC1 audience increased by almost 11 per cent (Deans, 2002b).

On 17 January, the share price of Granada, (one of the two largest ITV companies) fell by 4 per cent, as the impact of the audience figures hit the stock market (Barbed numbers snag ITV companies, 2002). There followed a period of continued turbulence in the industry, reflected in shock headlines in even the broadsheet press, such as '"5 million viewers lost" as TV ratings in chaos' (Davies, 2002), 'BARB's missing millions' (Day, 2002) and 'Discrepancies "make mockery" of TV ratings' (Plunkett, 2002). By June, Davies was still reporting that only 4,200 households were being surveyed (100 less than before the changeover) as well as concerns over how representative they really were. Because of this unusual focus on the validity of the organization's research, as the attention of the media continued to be directed at the organization, Day highlighted the inability of BARB's methodology to measure viewing away from the home: the timing of the 2002 World Cup matches was encouraging large audiences to gather around sets in more public venues, such as pubs and workplaces, and so the considerable extra viewing the event was generating was unaccounted for in the official industry audience figures.

Plunkett's analysis, however, highlighted a number of discrepancies which relate to the appropriateness of the new, 'improved' sample in representing the sub-populations that constitute the television regions. A controversial docudrama screened nationally on ITV1 on 9 July 2002 about the serial murderer Harold Shipman seemed from the research data to have been watched by 41 per cent of the audience in the Granada region, and almost a third less (28 per cent) in the Yorkshire region, even though Shipman's crimes were committed in both regions. Another ITV1 programme that week, *Everything Must Go*, received a 23 per cent share of viewing at the time nationally, but it achieved implausibly contrasting share figures in different regions: 63.6 per cent on Border Television and only 6.6 per cent on neighbouring Tyne Tees. When BARB Chief Executive, Caroline McDevitt, left the organization in late 2003, media reporting of her departure continued to link her with the crisis of January 2002 (Cozens, 2003).

Changing a sample *en bloc* as BARB did, exposes weaknesses in sampling as a means of producing reliable estimates and renders the methodology underpinning this type

of audience research very vulnerable to criticism. Sharp fluctuations in the results produced can only damage confidence in the validity of invisible audience measurement, which is why research organizations normally ensure changes to panels happen more gradually, often on a 'rolling' basis, to disguise inadequacies inherent in the orthodoxy of sampling. RAJAR 'rolls' samples forward in the majority of the TSA's it surveys, using completely new sets of respondents each quarter in only the largest—London, large regions and nationally. This has the effect of minimizing wide fluctuations in results. Otherwise, as in the case of BARB, changing samples wholesale, and so producing starkly differing results would too often cast doubt on the generalizability of findings among each sample.

Had the sample sizes, before and after BARB's changes, been much larger, it is likely that the discrepancies attributed to changing the sample in 2002 would have been less apparent. If it is accepted that larger samples can be more reliable than smaller ones, it must be because they are constituted with fewer constant errors built into them. What happened in 2002 was that any constant errors regularly produced by a mismatch between the first sample and the population it was supposed to represent were either removed in the change or replaced by other constant errors that, in turn, then began to produce markedly different results. The protests from broadcasters, that the new sample was disadvantaging them, inevitably claimed the new sample to be flawed, because it was producing different results to those to which they had been accustomed and on which their scheduling and budgetary decisions had been predicated. It remains possible, although not provable, that the new figures were more accurate than the old ones. Certainly, BARB argued that the initially low panel sizes were less of a shortfall than it appeared because the improvements in representativeness made the panel more effective per capita.

Part 5: Selecting Samples

Making changes to a sample in order to respond to naturally occurring changes to the demographics of the population, however, is clearly justified by the need to produce as valid a survey as possible. Table 4 shows how different audience research organizations set different parameters even for their largest populations, be they the whole of the UK or just Great Britain (excluding Northern Ireland). RAJAR and NRS also define 'adults' as being adults from 15 years of age, while GfK sets the same parameter at 16. The most authoritative data set from which they can derive their 'representative' samples is that produced every 10 years by the National Census (costing £255 million in 2001). This is, in theory, the complete enumeration of all units under observation, which is the only alternative to sampling in determining information about a population (Som, 1973, p. 12), but as such, it would be a prohibitively expensive undertaking for audience researchers. Even if it were affordable, if audience data were produced in this way, it would be far too infrequent to meet the needs of the industry. However, as the only actual enumeration of the population, the Census provides the industry with the closest indication possible of the location and size of each demographic group in the UK, and so it can be used to determine the strata used in

subsequent sampling. This survey, unlike audience research estimates, counts tangibles: real people, interviewed in person by a small army of Census workers who visit all the known permanent dwellings in the nation. Of course, such approaches are not infallible. In 1960 the US Bureau of the Census abandoned complete enumeration in favour of a 25 per cent sample of the US population, because the inaccuracies produced by inadequate enumerators in the previous Census had been so significant that the savings released by using fewer, better trained interviewers made the use of estimation according to a population to sample basis of 4:1 attractively cost effective (Som, 1973, p. 3). In the case of the UK Census, fallibility is due in part to random collection error of various kinds, and in part to non-co-operation by individuals, however, while never a hundred per cent accurate, producing more reliable demographic data about the whole population would be very difficult to achieve.

The audience research organizations establish 'representative' samples of the various populations within the Census by identifying individuals or households that closely match their demographics. For instance, BARB conducts its own annual Establishment Survey of 52,000 households to identify those which would be appropriately representative of each population, based on age, gender, regional and household size characteristics (BARB, 2003a). This survey is random, so BARB maintains every household in the UK has an equal chance of being selected, and depending on perceived deficiencies in the panel by comparison with the Census, new households may be recruited from the Establishment Survey in accordance with need, as existing panel members are removed.

The Census occurs only once in 10 years, so the release of its headline data can provoke audience researchers into making hasty adjustments of samples. In 2001 the Census found there were 900,000 fewer people living in England and Wales than were anticipated in the 1991 Census, 800,000 of them young men (Table 6). The Office for National Statistics considered migration and projection error rather than Census avoidance, to be the main reasons for the discrepancy (ONS, 2003).

Table 6 Percentage Change in Total Population Data from the 2001 National Census Results

Age group	Total individuals	All males	All females
0–4	−4.7%	−4.7%	−4.7%
5–9	−1.9%	−2.2%	−1.6%
10–14	−1.9%	−2.1%	−1.6%
15–24	−2.4%	−3.8%	−0.8%
25–34	−4.4%	−9.0%	+0.4%
35–44	−3.2%	−6.2%	0.0%
45–54	−1.4%	−2.2%	−0.6%
55–64	+3.9%	+3.9%	+3.9%
65 +	+3.4%	+3.3%	+3.5%
Total	−1.2%	−2.8%	+0.4%

Source: BARB, 2003c

Projections from 1991 had suggested population growth, rather than decline, and consequently some surveys had adjusted their estimates accordingly over the decade. In July 2003 RAJAR announced that a fall of 1.6 per cent in the population would be reflected in survey results from the second quarter of the year (RAJAR, 2003f). Consequently, stations were warned not to immediately interpret falls in listenership as sudden losses of listeners between quarters, but that they might simply reflect adjustments to the numbers of people in their TSAs. BARB was slower to react to the news, announcing in October 2003 the changes to its national and regional populations that would take effect from January 2004 (BARB, 2003c).

Although initially not due to any act of omission of their own, BARB and RAJAR (as well as NRS) were therefore routinely producing audience data that Census data now show to have been incorrect, in terms of the false premise on which they were based, for at least part of the 1990s and certainly more recently until 2003. Although the margin of constant error this discrepancy caused was not large, it was nevertheless concentrated in one particular demographic group: young men, who will have been over-represented in samples as a result. It follows that future projections, on which samples will be based, may not be totally reliable, either, and the release of the 2011 National Census data may well bring with it some more surprises.

Conclusion

The nature of sampling as a means of producing estimates of invisible consumption by audiences is one of an inexact science. The logic and the mathematics of sampling orthodoxy may be sound, but the practice of sampling in audience surveys seems to produce telling inconsistencies that demonstrably deny some audience research the legitimacy to which it aspires. The generalizability of such concerns is impossible to determine, but given the problematic nature of constructing representative samples which report accurately, the probability of most audience research producing accurate estimates of viewing and listening by the populations they are supposed to represent is not high.

The BARB experience of 2002, and the effect of constructing sample strata on the conclusions of census data that are tainted with 'correctional' projections of future trends, would suggest that the widespread use of stratified random sampling is not producing the kind of accuracy that audience research needs to be credible. As Stopher and Meyburg (1979, pp. 14–19) and Stuart (1962, pp. 10–13) observed, complete randomness eliminates selection bias of the sort that distinguished BARB's first panel from its second (both panels producing contrasting results). However the relative minutiae of detail required of surveys covering large numbers of sub-populations including multiple demographic, geographic and platform variables renders the survey size that would be necessary to reduce constant error below significant levels uneconomic. Stratification presents a degree of confidence in steering sample construction towards representativeness in smaller numbers, but as Yates warned, if done imperfectly, it can introduce an unacceptable incidence of constant error into the resultant survey data (1981, pp. 26, 264).

The challenges facing survey organizations wanting to produce accurate data are clear: to recruit respondents who are collectively truly representative of the populations they are supposed to represent, who do not behave in maverick ways, and who are all fully compliant with the various demands of data collection. As they can never know if they have achieved this, they, and the industries that use the data they produce, must continue to collude in presenting their findings as reliable: a practice made easier by producing consistent results rather than revealing wild fluctuations and by using permanent panels and rolling samples rather than performing regular purges of respondents.

Of course, the broadcasting and advertising industries could voluntarily refrain from their excesses in the use and publication of audience data in a number of ways: aggregating data over longer periods, only reporting on larger sub-groups, including more prominent 'health warnings' in publication and resisting the temptation to extrapolate from data collected in off-peak dayparts are just a few suggestions. Because radio and television audience research normally lacks controversy, the sampling behind its output is rarely scrutinized and few questions are asked about the representativeness of those who contribute to it. Being the sole currency of both production and airtime sales, it is the audience research that matters to producers and advertisers alike, and to the majority of producers, the possibility of its being an inaccurate measure of consumption is an unnecessary distraction from the business of achieving ratings success in the industry's only 'people market'.

Acknowledgements

Cultural Trends would like to thank Professor Andrew Crisell, University of Sunderland, Paul Allin, Office for National Statistics and Pam Hanley, research consultant, for their peer reviews.

References

Allin, P. (2003). Commentary 2 on: Radio audience research: challenging the 'gold standard'. *Cultural Trends, 45*, 85–91.

BARB (2003a). *Television measurement service.* London: Broadcasters' Audience Research Board Ltd. Retrieved October 10 2003 from http://www.barb.co.uk/about.cfm?flag=about

BARB (2003b). *Viewing summary.* London: Broadcasters' Audience Research Board Ltd. Retrieved October 30 2003 from http://www.barb.co.uk/viewingsummary.cfm?flag=viewingsummary

BARB (2003c). *Population changes for 2004.* London: Broadcasters' Audience Research Board Ltd. Retrieved October 10 2003 from http://www.barb.co.uknews.cfm?fullstory = true&newsid=97&flag=news

Barbed numbers snag ITV companies. (2002, January 18). *The Guardian.* Retrieved October 15 2003 from http://media.guardian.co.uk/broadcast/story/0,7493,635456,00.html

Beckett, A. (2001, November 20). Numbers game. *The Guardian, Media Guardian* suppl., 2.

Boon, P. (Ed.). (2003). *UK radio guide and directory.* Kettering: Goldcrest Publishing.

Collett, P., & Lamb, R. (1986). *Watching people watching television: Final report to the IBA.* Oxford: University of Oxford, Department of Experimental Psychology.

Cozens, C. (2003, July 17). McDevitt quits as BARB boss. *The Guardian.* Retrieved October 15 2003 from http://media.guardian.co.uk/broadcast/story/0,7493,1000326,00.html

Davies, A. (2002, June 20). '5m viewers lost' as TV ratings in chaos. *The Guardian.* Retrieved October 15 2003 from http://media.guardian.co.uk/broadcast/story/0,7493,740704,00.html

Day, J. (2002, June 20). BARB's missing millions. *The Guardian.* Retrieved October 15 2003 from http://media.guardian.co.uk/broadcast/story/0,7493,740765,00.html

Deacon, D., Pickering, M., Golding, P., & Murdock, G. (1999) *Researching communications.* London: Arnold.

Deans, J. (2002a, January 3). Ratings chaos leaves broadcasters in the dark. *The Guardian.* Retrieved October 15 2003 from http://media.guardian.co.uk/broadcast/story/0,7493,627201,00.html

Deans, J. (2002b, January 17). ITV falls 25% in new Barb ratings. *The Guardian.* Retrieved October 15 2003 from http://media.guardian.co.uk/broadcast/story/0,7493,634944,00.html

Electronic measurement ticks on. (2003, March 8). *Radio Magazine, 569*, pp. 3–4.

Emap (2003) *Big city network.* Retrieved October 14 2003 from emap.com/nav?page=emap.features&resource=435303

Frankel, L. (1969). The role of accuracy and precision of response in sample surveys. In W. Johnson & H. Smith Jnr (Eds.). *New developments in survey sampling.* London: Wiley.

Galaxy (2003). *Galaxy 105 advertising information.* Retrieved October 14 2003 from http://www.galaxy105.co.uk/listingsEntry.asp?ID=14811&PT=ContactInfo

GfK (2003, October 18). A broadcast media survey of Radio & TV audiences. *Radio Magazine, 601*, 7.

Gibson, O. (2002, January 4). TV ratings suspended for two weeks. *The Guardian*. Retrieved October 15 2003 from http://media.guardian.co.uk/broadcast/story/0,7493,627644,00.html

Goodway, N. (2003, October 23). Listeners switch off Capital. *Evening Standard.* Retrieved October 15 2003 from http://www.thisislondon.co.uk/news/business/articles/timid69514

Henry, G. (1990). *Practical sampling.* New York: Sage.

McDevitt, C. (2002). Summary of TV 2002 Conference Speech, Prague, 22nd March. London: BARB. Retrieved January 6 2004 from http://www.barb.co.uk/news.cfm?fullstory = true&newsid-53&flag=news

NRS (2003). *About NRS: Data available to subscribers.* London: National Readership Survey. Retrieved October 9 2003 from http://www.nrs.co.uk/open_access/open_aboutnrs/data_available/index.cfm

Office for National Statistics (2000). *UK 2000 Time Use Survey.* London: ONS. Retrieved October 14 2003 from http://www.statistics.gov.uk/timeuse/

O'Hara, J. (2003). The RAJAR survey. *Cultural Trends, 45*, 81–83.

ONS (2003). *Implications of the 2001 census results: why census shows fewer men.* London: Office for National Statistics. Retrieved October 16 2003 from http:www.statistics.gov.uk/census2001/implications.asp

Plunkett, J. (2002, July 11). Discrepancies 'make mockery' of TV ratings. *The Guardian.* Retrieved October 15 2003 from http://media.guardian.co.uk/broadcast/story/0,7493,753226,00.html

RAJAR (2002). *Quarterly summary, period ending June 2002.* London: Radio Joint Audience Research Ltd/IPSOS-RSL. Retrieved October 23 2003 from http://www.rajar.co.uk/QuarterlySummary/, 1 August.

RAJAR (2003a). *BBC World Service and three national commercial digital services join RAJAR survey.* London: Radio Joint Audience Research Ltd, 11 April. Retrieved October 30 2003 from http://www.rajar.co.uk/INDEX2.CFM?menuid=6

RAJAR (2003b). *About RAJAR: The RAJAR Research Service.* London: Radio Joint Audience Research Ltd. Retrieved October 15 2003 from http://www.rajar.co.uk/aboutshow2.cfm?aboutid=6

RAJAR (2003c). *RAJAR announces tests on second generation of electronic meters.* London: Radio Joint Audience Research Ltd, 15 October. Retrieved October 30 2003 from http://www.rajar.co.uk/INDEX2.CFM?menuid = 6

RAJAR (2003d). *Quarterly summary, period ending September 2003.* London: Radio Joint Audience Research Ltd/IPSOS-RSL. Retrieved October 30 2003 from http://www.rajar.co.uk/QuarterlySummary/, 31 July.

RAJAR (2003e). *Quarterly summary, period ending June 2003.* London: Radio Joint Audience Research Ltd/IPSOS-RSL. Retrieved October 14 2003 from http://www.rajar.co.uk/QuarterlySummary/, 31 July.

RAJAR (2003f). *RAJAR Bulletin, 54*, 17 July, London: Radio Joint Audience Research Ltd. Retrieved October 16 2003 from http://www.rajar.co.uk/INDEX2.CFM?menuid=9

Ratings, week ending 14 September. (2003, October 3). *Broadcast*, 46–47.

Reeve, N. (2002, April 13). Radio research needs updating! *Radio Magazine, 522*, 20.

Ritson, M. (2002, May 14). Are you paying attention? A new study shows that people spend most of their time avoiding TV ads. *Financial Times, Creative Business* suppl., 2.

Sabo, W. (2002, October 18). The portable people meter is your friend. *Radio & Records*, 9.

Som, R. K. (1973). *A manual of sampling techniques.* London: Heinemann.

Starkey, G. (2003). Radio audience research: challenging the 'gold standard'. *Cultural Trends, 45*, 43–68.

Stopher, P., & Meyburg, A. (1979) *Survey sampling and multivariate analysis for social scientists and engineers.* Massachusetts: Lexington.

Stuart, A. (1962). *Basic ideas of scientific sampling.* London: Griffin.

Thompson, M. E. (1997). *Theory of sample surveys.* London: Chapman & Hall.

Tryfos, P. (1996). *Sampling methods for applied research: text and cases.* New York: Wiley.

TV ratings are back. (2002, January 14). *The Guardian*. Retrieved October 15 2003 from http://media.guardian.co.uk/broadcast/story/0,7493,632616,00.html

War! (2002, June 15). *Radio Magazine, 531*, p. 3.

Wells, M. (2002, January 21). TV ratings system is a bad joke, says broadcaster. *The Guardian.* Retrieved October 15 2003 from http://media.guardian.co.uk/news/story/0,7541,636607,00.html

Yates, F. (1981). *Sampling methods for censuses and surveys.* London: Griffin.