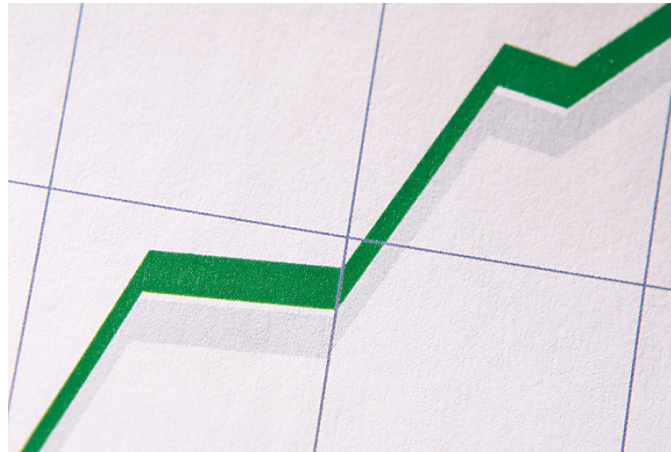


# Radio Audience Estimates



**What They Are,  
Where They Come From,  
And How To Use  
(And Not Use) Them**

## CHAPTER 1: THE BASIC ARBITRON METHOD

Prepared for

National Public Radio

January 2005

*James D. Peacock  
Peacock Research, Inc.*

# TABLE OF CONTENTS

Chapter 1: The Basic Arbitron Method.....	1
Table of Contents.....	2
Introduction.....	3
On Surveys and Estimates.....	3
On Forecasting.....	3
On Using The Right Tool.....	3
The Arbitron Method.....	4
Self-Reported Diaries And Their Limits.....	4
Implications Of The One Week Duration.....	5
Metros, TSAs, etc.....	5
Why "Quarter Hours".....	6
Other Methodological Strengths And Weaknesses.....	6

# INTRODUCTION

The NPR Research Department asked Peacock Research, Inc. to compile a guide to the nature of radio audience estimates commonly in use by public radio researchers. In particular, this guide is designed to build a bridge between the analysts who have to work with such data, and the end users who need to understand some of the strengths and weaknesses of the audience measures on which decisions are based.

This material will be available from NPR both as one single reference document and in extracts from the full report as a series of White Papers.

## On Surveys and Estimates

Radio audience estimates are *estimates*, first and foremost. Like anything else based on surveys, they are only approximations of what *may* have been happening in the total population.

Although radio audience estimates can be calculated to any number of decimal points, the reality is that sampling error often dwarfs the level of precision used when we present audience estimates. Even the best surveys can only approximate the behavior of a population simply because we haven't collected data from the entire population.

Furthermore, any audience estimation methodology has strengths and weaknesses. Any method is only as good as its ability to get survey participants to respond and to provide data that accurately reflects the behavior we think we're measuring.

To make these approximation challenges more difficult, we can only quantify *some* of the types of error that might affect the data we use. While we can estimate the range of error resulting from sampling alone (more on that later), we can never be entirely sure of the extent of bias that might be inherent in the methodology itself.

That doesn't mean the audience estimates are "bad," nor does it mean we can't make decisions based on them. But it does suggest that statistics should not prevail over common sense.

## On Forecasting

Not only do survey-based measures have significant limits, but radio audience estimates are also measures of *the past*. The numbers were based on a particular survey at a particular point in time, and they may or may not be good predictors of the future—or even of the present. The future often involves changes and factors which we haven't seen before, so even the best mathematical forecasts can be wrong.

Radio estimates are imperfect estimates of the past, no matter how much we may wish to have perfect, precise predictors of the future.

## On Using The Right Tool

Once we choose to live within the general limits of audience surveys, we still have other choices to make. There are many different ways to look at standard audience data, and each type of calculation brings its own caveats. Audience numbers that are meaningful in one con-

Radio audience estimates are  
imperfect estimates of the past.

But we tend to use them as if they  
were perfect and precise  
predictors of the future.

text can be misleading in another, and we'll try to sort out some of these challenges later in this report.

## THE ARBITRON METHOD

### Self-Reported Diaries And Their Limits

To understand the nature of the most common radio audience estimates, it's worth considering the characteristics of the Arbitron radio survey—the dominant source of radio ratings data in both commercial and public radio.

In a nutshell, an Arbitron ratings report reflects a series of one-week surveys of people aged 12 and older. Almost all Arbitron estimates reflect averages of multiple one-week surveys; the typical Arbitron estimate is an average of 12 of these one-week surveys.

At the heart of Arbitron surveys are the questionnaires used by survey participants to write down their radio listening—the infamous “diaries” pictured in Figure 1.

Within the diaries are pages for each day of a defined seven-day period (Figure 2).

Each potential respondent is recruited over the telephone from a reasonably comprehensive list of all possible telephone numbers (a variant of Random Digit Dialing). Those who agree (or more precisely, those who don't refuse) are mailed one diary per person 12+ in that household.

Each “diarykeeper” is expected to write down information about each time they hear<sup>1</sup> a radio for their designated survey period, which runs from a specific Thursday through the following Wednesday. Completed (and some incomplete) diaries are returned to Arbitron in the mail.

On average, Arbitron gets completed and usable diaries from about a third of the people they set out to survey. That “response rate” can be as low as 20-30% in the largest markets (which tend to be less cooperative in all surveys); cooperation is also lower than average for young adults and for Hispanic and African-American populations.

To compensate for differential rates of return, Arbitron weights its final sample through a method called sample balancing. This ensures that major demographic groups contribute to the ratings in proportion to their population size.

Low response rates are probably the greatest weakness of the Arbitron diary method. Unfortunately, we don't know very much about how those response rates affect the reported audience estimates. We know that the two-thirds of the population which doesn't cooperate probably has different radio listening than those who do participate, but there's little useful re-

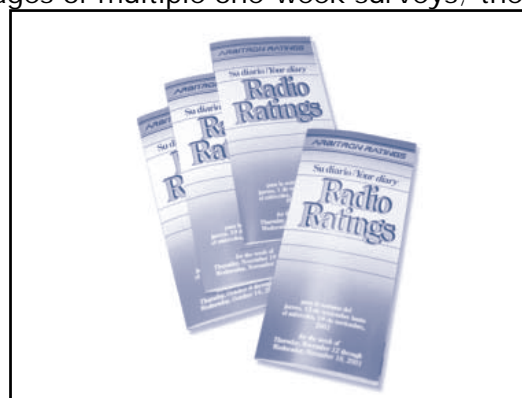


Figure 1: Arbitron Diaries

A sample diary page for Thursday. The page is titled 'THURSDAY' and contains a grid for recording radio listening. The grid has columns for 'Time', 'Station', and 'Place'. The 'Time' column is divided into 'Early Morning (6:00-9:00 AM)', 'Midday (9:00 AM-12:00 PM)', 'Late Afternoon (12:00 PM-5:00 PM)', and 'Night (5:00 PM-11:00 PM)'. The 'Station' column is divided into 'Full-Service', 'Adult Contemporary', 'Country', 'Classical', 'News/Talk', 'Sports', and 'Other'. The 'Place' column is divided into 'In Car', 'In Home', 'In Office', 'In Store', 'In Restaurant', 'In Public Place', and 'Other'. The grid contains several entries with 'X' marks indicating listening. For example, in the 'Early Morning' section, there are entries for 6:00-7:15 on KSTU, 7:15-7:40 on KMRB on the air, and 7:40-8:30 on WDR. In the 'Midday' section, there is an entry for 3:00-4:00 on KSTU. In the 'Late Afternoon' section, there is an entry for 4:20-11:25 on Air Country Show. In the 'Night' section, there are entries for 7:00-8:00 on KMRB and 11:30-12:15 on Super Game Game. At the bottom of the page, there is a checkbox and the text 'If you didn't hear a radio today, please mark it here.'.

Figure 2: Diary Page

<sup>1</sup> Note that Arbitron uses the word “hear” instead of “listen.” This is a conscious effort to capture *all* radio exposures, not just those for which the respondent was actively listening.

bly has different radio listening than those who do participate, but there's little useful research on just *how* different they are as radio listeners.<sup>2</sup>

Despite that limitation, however, the Arbitron diary method has been accepted by industry researchers as the best available method for collecting seven days of radio listening behavior from consumers. Other methods might be better if we needed to collect less listening information, but for current industry needs, the seven-day diary has survived the test of time as the best affordable method.

## Implications Of The One Week Duration

One key component of the Arbitron method is that participants provide only one week of data. While that's fairly demanding of the participants, it does impose some limitations on users of the survey data.

In particular, Arbitron data can never tell us directly how many different people are reached by a particular station or network over an extended period of time. Arbitron can tell us how many different people are reached by the station or network in an average *week*. But we can only estimate roughly how many different people might be reached over a longer interval (e.g., a month or a year).

There are several ways to make such an estimate, especially for a one-month period. But all of those methods use mathematical models to derive longer-duration estimates.

What's important to remember is that Arbitron audience estimates of "number of different people reached" (detailed definitions will follow) are for the *average week*. We don't know, and Arbitron can't tell us directly, how many different people are reached during an average month, or an average *year*. We know that those numbers will be larger than the Arbitron estimates of weekly reach, but there's no direct measure of those longer-term audiences—at least, not yet. (See the discussion in a later White Paper of Arbitron's efforts to develop a Portable People Meter which could generate reach estimates for periods longer than a week.)

## Metros, TSAs, etc.

It's also important to remember that all Arbitron estimates of audience are linked to a particular geography. A typical Arbitron audience estimate is based on an Arbitron-defined "Metro." These Metros are often similar or identical to U.S.-government defined Metropolitan Statistical Areas (MSAs), but the number of exact matches is only slightly over half of the markets. Many Arbitron markets comprise unique aggregations of geographies that may not align perfectly with other definitions of that "market." That's important for several reasons.

First, users sometimes want to use "market data" (e.g., Census-provided income averages) in combination with Arbitron data to understand their service areas. But users have to be careful that the geographic definitions of those two different data sources are truly comparable.

Second: The actual values of many audience estimates (e.g., ratings and projections, defined below) are directly related to the geographic definition being used. Almost all numbers reported as percentages by Arbitron have a geographically-defined "total" as the numerator of the percentage. For example, a thousand listeners divided by 100,000 in the Metro will be a different "rating" than, say, a thousand listeners divided by 10,000 in a county.

And third: These reporting geographies used by Arbitron may or may not link closely to the actual coverage area of any given radio signal. A station whose signal only covers half of the Arbitron-defined Metro will be heavily affected by the fact that its signal only reaches half of what Arbitron considers to be "the market."

---

<sup>2</sup> It's widely believed that diary responders listen to more radio than do the nonresponders, and that has some intuitive appeal. But past research has suggested that such patterns may not hold for all demographics.

These are not flaws in the Arbitron method. But geography is an important caveat for those looking at Arbitron data.

Arbitron does provide data based on other geographies, including Total Survey Area (TSA), Designated Market Area (DMA, a TV-market definition), and even national-level estimates for networks. But Metro estimates are by far the most commonly used Arbitron data.

## Why “Quarter Hours”?

Most Arbitron data are based on something called a “quarter-hour.” Although diarykeepers can write down any start and stop times they want for each listening event, Arbitron always translates those diarykeeper entries into units of 15 minutes. More specifically, Arbitron credits listening to specific blocks of time within each hour—the quarter-hours from :00 to :14, from :15 to :29, from :30 to :44, and the block from :45 to :59. If a diarykeeper indicates listening for at least five minutes in any one of those quarter-hour blocks, the entire quarter-hour is counted as listening by Arbitron. Conversely, if a diarykeeper writes down only four minutes of listening within one of those blocks, no listening is credited for that quarter-hour.

This peculiar convention goes back many, many years, and few current researchers know of its exact origins. But it was probably created to account for the expected imprecision of diarykeepers, an acknowledgement that diarykeepers may not be all that exact in writing down their start and stop times.<sup>3</sup>

Whatever the origins, however, users need to be aware of this aspect of Arbitron data. In particular, we need to remember that processed Arbitron data don't tell us exactly when any one respondent started or stopped listening; we only know that they reported (or were credited with) at least five minutes of listening within a particular quarter-hour. For example, unless we go to a lot of difficulty in examining raw data, we won't know whether a particular brief program element caused them to tune away at a particular moment. We also won't know whether they went straight to another station when they tuned out the prior station; there could have been up to ten minutes of other activity in between what appears to be continuous listening.

The bottom line: Don't confuse apparently-precise start and stop times in processed Arbitron data with actual knowledge of when a person began or ended their listening session.

## Other Methodological Strengths And Weaknesses

For the record, Arbitron data users need to be aware of a few other caveats concerning the Arbitron method:

- Respondent ambiguity: Arbitron users who take the time to review the actual diaries tabulated for a particular survey are often surprised at the extent to which Arbitron's professional judgment plays a role. Diarykeeper entries can be ambiguous (e.g., incomplete duration data, or unclear station identification), and Arbitron has to interpret the intentions of the respondent. The process is well defined and controlled, but the fact is that interpretation is a frequent reality with written diary entries.
- Operational mistakes: Arbitron's operations are about as good as such a large-scale process can be, but even Arbitron makes mistakes. If something seems wildly implausible in the audience estimates, it pays to have Arbitron double-check their processes.
- Station mistakes: In some areas, Arbitron is only as good as the data provided by radio stations. If a station provides wrong information about itself or its programming, some of the listening may be misattributed by Arbitron. (Of course, it's also

---

<sup>3</sup> And in fact, Arbitron has done manual tallies that show how diarykeepers themselves tend to “round off” their entries to the top and bottom of the hour much of the time.

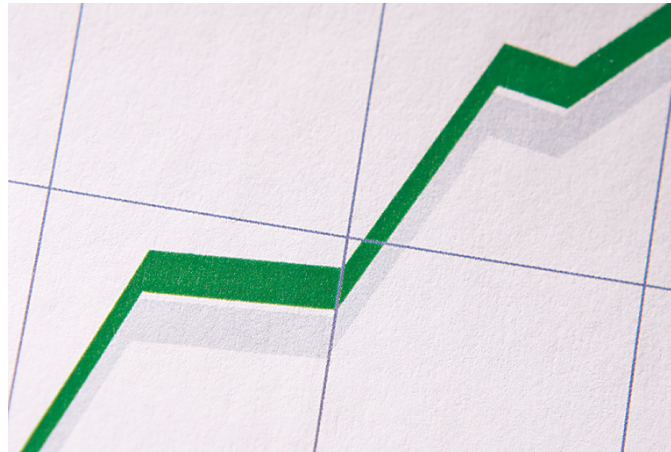
possible that a station provided false information knowingly, or that a station employee tried to affect the ratings improperly. Those situations are rare, but they do occur, and they can be hard to identify.)

- Overly-precise data reporting: Under pressure from subscribers, Arbitron often reports data to a greater degree of precision (e.g., decimal points) than the underlying data can really support. We'll discuss this further in a later section on Reliability.

Overall, audience data from Arbitron are solid estimates of radio listening. The method is executed well, and so far, the industry has shown little willingness to pay for more expensive approaches. But the data do have limits, and users need to remember to add reasonable judgment to their analyses. Audience estimates are just that—estimates.



# Radio Audience Estimates



**What They Are,  
Where They Come From,  
And How To Use  
(And Not Use) Them**

## CHAPTER 2: TYPES OF ARBITRON DATA

Prepared for

National Public Radio

January 2005

*James D. Peacock  
Peacock Research, Inc.*

# TABLE OF CONTENTS

Chapter 2: Types of Arbitron Data .....	1
Table of Contents .....	2
The Basic Arbitron Data .....	3
What The Respondents Provide .....	3
Three Kinds of Numbers .....	3
Two Components of "Listening" .....	4
AQH, An Overall Measure .....	4
Applications and Limits .....	5
AQH Pros and Cons .....	5
Cume Pros and Cons .....	6
TSL Pros and Cons .....	6
Differences Among Ratings, Shares, and Persons (Projections) .....	6
The Issue of Daypart Variation .....	7

In this second chapter, we'll discuss the nuts and bolts of the basic Arbitron radio listening data, including their applications and limitations.

## THE BASIC ARBITRON DATA

### What The Respondents Provide

In Chapter 1 of this series, we discussed the survey mechanics of Arbitron radio surveys. Now let's focus more carefully on the Arbitron radio data itself. While the analysis of audience data can seem incredibly complex, the underlying data are actually very basic. In addition to limited demographic and socioeconomic data about themselves, Arbitron diarykeepers provide the following information about each listening event recorded in their diaries:<sup>1</sup>

- Start time
- Stop time
- Station identification ("Call letters, dial setting, or station name" and "AM or FM")
- Location ("At Home/In A Car/At Work/Some Other Place")

That's it—time, station, and location. From that basic information, a myriad of summary measures are derived.

### Three Kinds of Numbers

From the raw data provided by diarykeepers, Arbitron computes three basic types of summarized numbers—ratings, shares, and persons estimates (or projections).

**Ratings:** By strict industry definition, a rating is particular kind of audience estimate—a percentage for which the denominator is the total population of a demographic group and a particular geography. Therefore, a rating of 5.0 (or 5%) represents 5% of a particular population.

**Shares:** Though a share is also expressed as a percentage, the denominator is different from that of a rating. For a share, the denominator is the number of people who are *listening* to radio among that demographic and geographic group.

**Persons Estimates (Projections):** This type of estimate is an actual count of estimated listeners, rather than a percentage. A persons estimate forms the numerator of ratings and shares, but is also reported separately (usually in multiples of 100).

Each of these three types of estimates is usually defined by specific demographics (or other respondent characteristics), a specific geography, and a specific block of time (e.g., a "daypart" like Morning Drive).

#### Diary Entry Examples

Call Letters only							
3 : 30	4 : 00	WBAL		x			x
4 : 30	5 : 30	WPOC			x	x	
All entries containing Call Letters							
6 : 00	7 : 15	WKYS 93.9			x	x	
Frequency only							
11 : 30	11 : 45	103.1			x		x
All entries containing a Frequency							
6 : 00	6 : 20	88.1 NPR			x	x	
Station Name only							
5 : 15	5 : 30	The Bay			x		x
5 : 30	9 : 30	ESPN Radio		x			x
All entries containing a Station Name							
3 : 10	4 : 00	Heaven 600		x			x
All entries containing Programming							
7 : 10	7 : 30	105.7		x			x
:	:	Howard Stern					

<sup>1</sup> With editing from Arbitron if the data are incomplete or ambiguous, as noted in White Paper #1.

Here's an example of how the same audience data could be expressed as each of these

Arbitron reports that:

- WAAA had 10,000 listeners at a particular point in time.
- Meanwhile, there were 100,000 people listening to any radio station at that time.
- And there were 500,000 people living in that geography.

WAAA's audience estimate could then be expressed as follows:

- The **persons estimate** (or projection) equals 10,000, which would usually be expressed in hundreds as a value of 100.
- The **share** would equal those 10,000 listeners divided by the 100,000 people listening to radio:  $10,000 \div 100,000$  as a percentage = 10.0%, or a share of 10.0.
- The **rating** would equal those 10,000 listeners divided by the 500,000 people in the corresponding population:  $10,000 \div 500,000$  as a percentage = 2.0%, or a rating of 2.0.

three measures:

Other types of measures are possible, but these are three "Big Three" reported and used most often.

## Two Components of "Listening"

In addition to the three types of numbers above, there are two basic kinds of listening data used in such calculations. As you'll see, these have their roots in the kinds of data provided by respondents.

**Cume (or Reach):** This measure is a simple "Yes or No" indicator of whether someone listened or not. As defined by Arbitron, a person is counted as part of a Cume audience if they listened for at least five minutes within a particular time period. A Cume audience can be expressed as ratings or as persons estimates. In theory, one could also compute a Cume *share*, but for reasons we won't cover here, that particular measure often isn't practical to calculate.

**Time Spent Listening:** While cume is a count of people in the audience, Time Spent Listening (or TSL) is an estimate of how *much* each listener listened during a particular time period. TSL is usually reported in either Quarter Hours or in hours and minutes.

Those two building blocks tell us *how many* and *how long*, from which we can compute the most frequently used of radio audience estimates...

## AQH, An Overall Measure

**Average Quarter Hour:** By far, the most common Arbitron numbers are those in the form of an Average Quarter Hour, or AQH, estimate. An AQH estimate is essentially the combination of Cume and TSL—of the number of people who listened at all, weighted by the extent of their listening.

The primacy of AQH has its roots in advertising. AQH numbers are the closest thing in radio to a "commercial audience"—an estimate of the number of people who were exposed to a single radio commercial. Here's why...

An Average Quarter Hour audience estimate is an approximation of the number of people who were in the audience *at an average point in time*. By considering how many people listened

at all (Cume) and the length of their listening (TSL), we can estimate how many people were listening at an average moment during that time period.

For an advertiser, this gives a reasonable proxy for commercial audience—for the number of people likely to have been exposed to a single commercial during the time of day (daypart) in question.

In fact, this is a fairly crude proxy for momentary audience because of the rules for credit-ing listening in quarter-hour chunks that were discussed in chapter 1.

And if we're looking at a broad daypart, it's even cruder, since the number of listeners may vary considerably over the hours within the daypart. It may be accurate as an average, but the average may not fit individual moments especially well.

But this is as good as it gets with diary measurement.

**Flavors of AQH:** Average Quarter Hour measures come in all three types described earlier (ratings, shares, and persons estimates). An AQH persons estimate is the number of persons estimated to be listening *during an average quarter hour* for a particular combination of station, daypart, demographic, and geography.

An AQH rating is the estimate of AQH persons expressed as a percentage of the correspond-ing *population*. And an AQH share is the AQH persons estimate expressed as a percentage of the corresponding number of people *listening to any radio on an AQH basis*.

Beyond its appeal to advertisers, the AQH family of estimates has clear virtue as a unifying measure of station, program, or radio listening. In one number, it encapsulates the number of listeners and the extent of their listening.

Furthermore, AQH has some mathematical attributes that enhance its utility. In particular, AQH audiences can be added across stations; cume listeners cannot (at least not as easily, since a person can listen to more than one station, introducing double-counting).

## APPLICATIONS AND LIMITS

Now that we've defined the building blocks of radio audience measurement, let's consider the strengths and limitations of these measures.

### AQH Pros and Cons

By definition, AQH audiences are smaller (and probably more humbling...) than Cume audi-ences. That's because a Cume listener only has to listen for a very brief interval (five minutes) to be counted as a listener. But an AQH listener, in a sense, has to count all over again for each quarter-hour during a time period; as soon as they stop listening, they stop contributing to the AQH estimate.

For an advertiser or sponsor, an AQH estimate is the most appropriate one to use since it considers the actual probability of an exposure happening at any one moment during a time pe-riod. It can have similar benefits for planning internal promotions where we want to know how many people are likely to be reached by any given announcement.

It also has considerable use for programmers, as it provides a single measure that combines the effects of reaching people and of getting them to listen longer.

The downsides of AQH measures mostly involve the definition of the daypart. As mentioned above, an AQH measure is an average; like most averages, it can mask variations within its computation. For example, a Morning Drive AQH number is a good average, but it really doesn't



tell you much about variations *within* Morning Drive. A decent AQH Morning Drive audience could in fact include very poor performance in one hour being masked by excellent performance in a different hour within that daypart. Therefore, the longer the daypart that defines an AQH estimate, the less likely that estimate is to accurately reflect any given hour within that daypart.

Finally: While an AQH measure is a good overall measure, it can be achieved several different ways. A particular AQH value could represent a lot of people listening for brief periods; it could also represent a few people listening for longer periods. Programming and planning conclusions need to consider the component parts of an AQH measure to determine the dynamics of the audience.

## Cume Pros and Cons

Cume audiences are the largest (and thus, the most gratifying) numbers available from Arbitron. But there are reasons other than size to find utility in Cume numbers.

Fundamentally, audience growth can occur in two ways—getting more people to listen, and getting listeners to listen longer. Cume audience estimates help track the former.

The problem is that Cume and Time Spent Listening can be negatively correlated with each other. If you set out to get more new listeners, it's likely that your Cume will go up while your average TSL goes down. That's because newer listeners may only be sampling portions of the station's or network's offerings.

So Cume audience estimates are useful, primarily for tracking the number of people that sample a station or network. But for an increase in cume to be truly beneficial, it should correlate with an increase in AQH; otherwise, the growth in new listeners had a price in disproportionate loss in TSL. Only if AQH goes up does a growth in cume really spell long-term success, at least in a majority of cases.

## TSL Pros and Cons

As with Cume audience estimates, Time Spent Listening estimates can shed more light on changes in a station's or network's audience. But if viewed in isolation, it too can be deceiving. An increase in TSL is good only if there isn't a high price in Cume; and a decrease in TSL isn't necessarily bad if it was driven by an increase in Cume.

In the end, TSL and Cume need to be evaluated together, and with an eye on AQH for overall effects.

Furthermore, as we'll discuss in a later White Paper, TSL is particularly prone to reliability (stability) problems, at least relative to AQH and Cume measures. A TSL estimate is computed only on a base of a station's own listeners, and that reduced sample size makes it one of the less reliable numbers that can be computed from Arbitron data. Changes in TSL need to be relatively large before being considered "real" (or actionable).

## Differences Among Ratings, Shares, and Persons (Projections)

Finally, we should touch on the unique characteristics of ratings, shares and persons estimates. Across these statistics, it's the share which stands out, while ratings and persons estimates have much in common.

Ratings and persons estimates are absolute measures, in that they express a station's audience in terms of a specific count of people, or in terms of a percentage of the population. Other stations' audiences are not directly related to these two kinds of computations.

Ratings and persons estimates also have some statistical properties that makes it easier to estimate the amount of sampling error affecting such estimates.

Shares, however, are relative; they express one station's audience estimate relative to the survey estimate of "all listening." For example, if 5% of the population listens to WAAA, and 20% of the population listens to any radio, then WAAA would have a 25% share of listening (5/20).

That's why shares are always numerically larger than their corresponding ratings; the numerators are the same, but shares are based on a much smaller denominator.

That also introduces a loss of reliability (i.e., it causes an increase in sampling error), since the base of a share percentage is about one fifth of the base of a rating. There are other statistical complexities with shares, too, and even Arbitron doesn't attempt to estimate the sampling error around its published shares.

Those are some of the reasons this author tends to favor using ratings and persons estimates over shares for most applications.

However, there's at least one situation in which shares can have greater intuitive utility. It's common for broadcasters to think of a market as having a total *potential* audience for a particular format or type of programming. For example, it's useful to consider the total audience for a particular type of format in a market, regardless of the number of stations. That multi-station estimate can be a relatively stable number, and it's reasonable to think of that total audience as the theoretical maximum for a station in a particular format.<sup>2</sup>

Under those circumstances, it can be useful to assess a single station's share of that format potential. For example, if the total audience across all news-oriented stations equals a 10.0 rating, and a particular station has a 4.0 rating, the station's share of the format is 4/10, or 40%. That kind of intelligence is a useful way to think of market potential.

But otherwise, this author recommends focusing on AQH ratings and persons estimates rather than shares.

## The Issue of Daypart Variation

Another caution about many measures has to do with expectations. Arbitron numbers are often used to compare programs to each other, or stations, or announcers, and so on. But even when it's acceptable to compare the two numbers mathematically, we still need to make sure that the comparison is *grounded in appropriate expectations*.

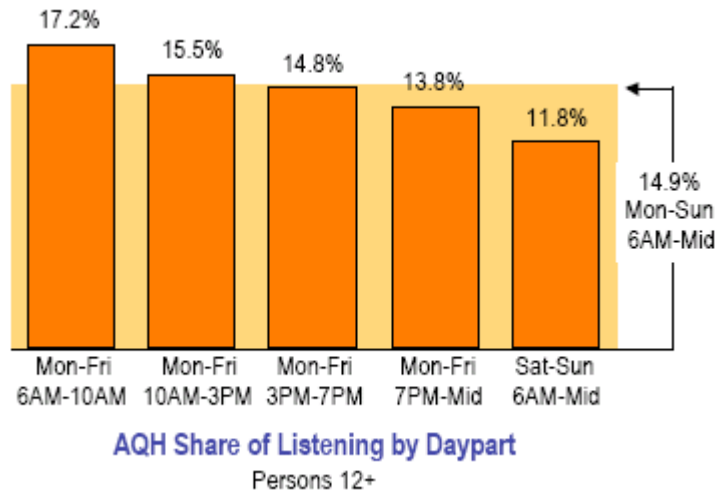
In a broad sense, not everything that affects an audience estimate is under a programmer's control. For example, a station whose signal only covers half of an Arbitron Metro should have different expectations about the size of its audiences than does a station with full or high-power coverage.

Analysts usually remember to account for such exogenous factors, but there's one that is often overlooked—the tendency of listeners to have different patterns and preferences at different times of the day. For example, it's simply a fact of life that news and talk tend to do better in so-called Drivetimes, especially in Morning Drive (Monday-Friday, 6AM-10AM).

Here's Arbitron's summary of how well commercial News/Talk/Information formats do by time of day (from Arbitron's publication, *Radio Today: 2003 Edition*):

---

<sup>2</sup> NPR Research is conducting some new analysis to determine whether a "format share" in a market is affected by the number of stations in that format. That makes intuitive sense, but we'll hold off further discussion of that possibility until we can quantify it better.



**Figure 3: Commercial News/Talk Stations**

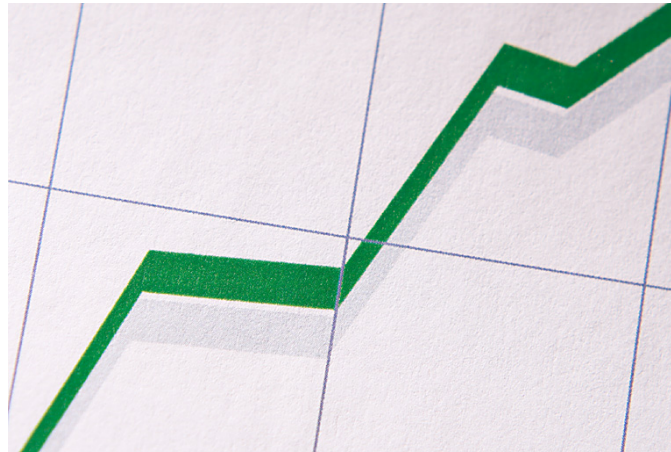
As you can see, commercial News/Talk/Information stations across the country tend to have 15% stronger AQH shares in Morning Drive than they do overall. Not surprisingly, the Evening and Weekend shares are smaller than the average.<sup>3</sup>

Of course, there are exceptions for some stations and programs. But in general, it's fair to say that our *expectations* of any given format should vary by time of day. The fact that a News/Talk station has lower ratings or shares in dayparts other than Morning Drive should be considered as much of an external factor as signal strength, and should not be ignored in making comparisons across dayparts, or from one daypart to a station's average.

Keep this issue in mind as we consider supplemental Arbitron-derived estimates, too. That's the subject of a future chapter.

<sup>3</sup> This chart focuses on shares rather than ratings to control for the varying amount of total radio listening by daypart. The same chart based on *ratings* would show even more dramatic differences by time of day.

# Radio Audience Estimates



**What They Are,  
Where They Come From,  
And How To Use  
(And Not Use) Them**

## CHAPTER 3: LISTENER CHARACTERISTICS

Prepared for

National Public Radio

November 2004

*James D. Peacock  
Peacock Research, Inc.*

# TABLE OF CONTENTS

Chapter 3: Listener Characteristics.....	1
Table of Contents .....	2
Caveats About Demographics .....	3
Race and Hispanic Origin .....	3
Language Usage .....	3
Income and Education.....	4
Supplements to Arbitron.....	5
Recontact Studies .....	5
Imputation, Fusion, And Other Appendages .....	6
Coming Next.....	6
About The Author .....	7

This is the third in a series of six NPR White Papers on the subject of understanding radio audience estimates. In earlier papers, we discussed how Arbitron collects its basic radio listening data, and how those elements are computed into the most common audience estimates. In this paper, we'll talk about an equally important part of radio data collection—the classification of diarykeepers into groups.

## CAVEATS ABOUT DEMOGRAPHICS

To complete our discussion of Arbitron-based data, we need to consider the non-radio data used in tabulations. The Arbitron process collects limited data about the diarykeepers themselves, and these classification variables have some limitations, especially within the categories below.



### Race and Hispanic Origin

A decade or two ago, it seemed relatively simple to classify someone by race and by country of origin. More recently, Census research has shown us that both variables are much more complex than we used to think, but Arbitron still resides in the land of simplicity.

For now, at least, Arbitron still acts as if a person can only be one race. Similarly, their procedures still treat race and Hispanic origin as a single variable; if a person is African-American, they cannot simultaneously be Hispanic (an obvious problem for many people of Puerto Rican origin). For now, Arbitron persists in using a single classification scheme that includes (in simplified form) only the choices Black, White, Asian, or Hispanic.<sup>1</sup>

Furthermore, at present, Arbitron classifies all residents of a household the same way. The person who agrees to the survey for their household is asked whether they consider the *household* to be Black, White, Asian, or Hispanic. This procedure is related in part to Arbitron's need to vary survey procedures at the household level (e.g., the set of Differential Survey Treatments used to encourage minority responses).

To Arbitron's credit, they go to some lengths to assure that the population data they use for weighting targets represent the same category scheme. Furthermore, as this is being written, Arbitron is conducting a test of procedures that would permit someone to choose multiple races, to finally separate Hispanic origin from race, and to classify at the personal level.

But until that test is complete and new procedures are implemented, users need to remember what's actually meant by Arbitron's classifications. "Blacks" are actually non-Hispanic Blacks, and the race/ethnicity classification is done at the whole-household level. Neither is a good reflection of today's reality, and we can only hope that the new test is successful.

### Language Usage

Arbitron has been under some pressure to approach the measurement of Hispanics in a different way. One of the strongest predictors of radio station listening is which language is preferred by the listener—especially whether or not someone is dependent on the Spanish language. While the classification of "Hispanic" ethnicity is a useful surrogate and can be shown to correlate with radio listening, the reality is that language preference is a better way to classify in this domain.

---

<sup>1</sup>. In fact, Arbitron does use two questions—one for Hispanic origin and one for race. But if someone indicates the answer Hispanic, the race question is not asked.

However, language preference is notoriously difficult to measure accurately and reliably. The Nielsen TV ratings firm has published a great deal of research in this area, and we've learned that:

- There are many different ways to segment language usage;
- There are many different ways to define "preference";
- The answer someone gives today may not match what they tell you tomorrow;
- The language used by an interviewer may have some effect on the answers;
- The ability of one person in the household to properly classify the language preference of another resident is flawed;
- And there are no Census-based estimates of language usage that are useful for media research.

Up to this point, Arbitron has limited its activity in this area to simply classifying its diarykeepers by a simple language preference question during the first telephone call, asked about each person in Hispanic households:

"Thinking about the languages (you/he/she) use(s) in the home, would you say (you/he/she) speak(s)..."

- ONLY SPANISH in the home,
- mostly Spanish but some English,
- mostly English but some Spanish,
- or ONLY ENGLISH in the home?"
- DO NOT READ: Both Equally"

Diarykeepers are classified as Spanish Primary if the response is "only Spanish" or "mostly Spanish."

Arbitron's current approach allows analysts to segment listening by language preference. But it does not as yet deal with whether Spanish Primary diarykeepers are properly represented in Arbitron surveys. To deal with that issue, Arbitron needs to have universe (population) estimates to which its surveys could be weighted.

Arbitron has announced that it plans to acquire language-related universe estimates from Nielsen, since Nielsen now conducts its own "language enumeration surveys" in most major markets. Overall, this is likely to be a good change for Arbitron, but the modification awaits changes in Arbitron's software systems that won't be complete until 2006. It will also need careful scrutiny by industry researchers on the details of its implementation.

For now, users need to understand that Arbitron data based on classifying respondents by language should not be compared to other sources of language-use data. The Arbitron process is unique and of Arbitron design. Furthermore, users need to remember that proper representation of Spanish-dependent Hispanics is not fully assured by current Arbitron procedures.

## Income and Education

Most researchers are aware that Income and Education are difficult to measure. Not all consumers are willing to answer such questions, and those that do aren't always honest or accurate.

Arbitron data are no better or worse than other sources in this regard (except possibly compared to the extensive efforts of government surveys). But remember that Arbitron tabulations by income and education include a nontrivial number of people who don't answer these questions, and that can throw off the reported distribution of audiences. Furthermore, there's good reason to believe that consumers aren't as accurate in their reporting of income and education to survey researchers like Arbitron as they are to Census and related interviewers.

Overall, don't consider Arbitron income and education data as "gospel." In particular, there's dilution of such data by nonresponders and by people who aren't really in their reported category.

## SUPPLEMENTS TO ARBITRON

To enhance the usage of Arbitron data, researchers sometimes supplement Arbitron with data from other sources. In particular, users have several tools at their disposal to add additional attributes to the Arbitron data.

### Recontact Studies

One of the most common ways to enrich Arbitron survey results is through a "recontact study." For example, the report on public radio called *Audience 98* by the company Audience Research Analysis (ARA) was based in large part on this type of supplemental survey.

During the original Arbitron survey, the typical respondent provides only a limited amount of information beyond their radio listening—age, gender, race, and a few other common characteristics. To help users go beyond this limited data, Arbitron makes it possible for subscribers to query those diarykeepers a second time to collect additional data. ARA, for example, surveyed past diarykeepers about many of their public broadcasting-related behaviors (donations, etc.).

The results of the second ("recontact") survey are merged with the original Arbitron data for those diarykeepers into a single database. That set of data can then be used to relate radio listening patterns with other kinds of respondent attributes.

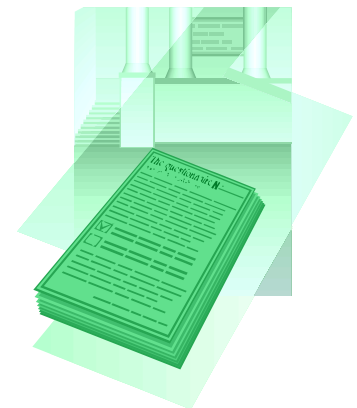
This can be a very useful tool for detailed analysis of radio listening. And in fact, it's sometimes the only practical way to combine high quality radio data with detailed demographics, socioeconomic indicators, lifestyle indicators, etc.

But this method has significant limitations that need active consideration. Perhaps most important is the uncertain effect of a low net response rate for the unified database. The original Arbitron survey would have a response rate of about 30-35%, and that's the starting place for the second survey. The second survey will have its own problems with cooperation, and the net response for people who provide data twice can be significantly below Arbitron's rate.

For example: In *Audience 98*, ARA mailed recontact surveys to 15,000 Arbitron diarykeepers, but only about 8,000 returned their questionnaires. That means that the *Audience 98* response rate would be only about half (8,000/15,000, or 53%) of Arbitron's. Using current Arbitron response rate performance, that would suggest that a recontact survey would have a net response rate of only 17-18% overall (and lower for large markets and tough demographics).

The similar *Public Radio Tracking Study* (1999) reports a recontact rate of 66%—an estimated net response rate of about 20% using Arbitron's current survey metrics.

In addition, recontact studies pose particular problems for tabulation. In the original Arbitron survey (and in most large scale surveys), a process known as sample balancing (or



“weighting”) is used to compensate for imperfect demographic representation in the survey. For example, surveys often end up with too few Men 18-24 compared to the population; the sample balancing/weighting process compensates by having the Men 18-24 who did participate count for more in the tabulations.

In principle, a recontact study has the same challenge, and it should use the same solution. At a minimum, a recontact study would benefit from weighting the final respondents back to the weighted distribution of the respondents who were selected from the Arbitron database.

Fortunately, the major published studies relied on by public radio (*Audience 98*, the *Tracking Studies*, etc.) have accounted for that potential source of bias through a careful re-weighting by age, race and ethnicity. Users of recontact studies should insist on such procedures for any other such projects.

Even with re-weighting, though, the original problem of double nonresponse remains, and it exists for some types of respondents more than others. When a survey only measures one out of five or six people in the population, users shouldn’t attribute too much precision to the outcomes. Informed judgment becomes more important than ever, and the researcher should provide full information to the user about the actual net response rate of the study.

## Imputation, Fusion, And Other Appendages

In addition to conducting a second survey with Arbitron diarykeepers, there are other less direct means of attributing extra characteristics to those respondents.

One of the most common ways to estimate additional characteristics of Arbitron diarykeepers is through geographic imputation. Systems like the Claritas Corporation’s PRIZM system are frequently used to impute economic characteristics to addresses based on the characteristics of the neighborhood of each diarykeeper household. Under the heading of “birds of a feather...”, a PRIZM-based analysis assumes that each household in a neighborhood has the average characteristics of its neighbors. The definition of “neighborhood” can range from Census block group on up to Zip code.

While this system is indirect—we don’t actually know the precise characteristics of each household—it avoids the disadvantages of additional data collection. We are only approximating the socioeconomic characteristics of the households in the Arbitron database, but at least we don’t lose any of those households through additional nonresponse.

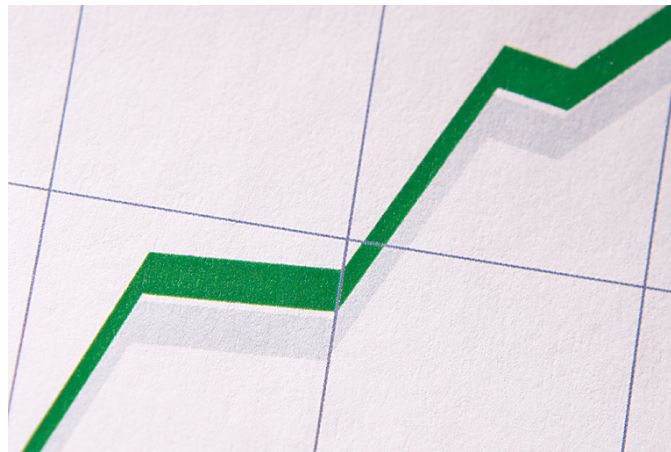
In principle, there are many ways to impute additional characteristics from one database to another. For example, the advertising industry periodically flirts with a statistical technique called “fusion” whose purpose is to impute characteristics from one survey database to another. Fusion has been shown to be a reasonable surrogate for direct measurement under certain circumstances (though not all).

The key limitation of almost all imputation methods is what’s called “attenuation.” Typically, a tabulation based on imputed characteristics will show less variation (i.e., will display a narrower range of values) than will direct measurement. For example, if a station has twice the incidence of high-income listeners as other stations in reality, a tabulation from imputed data will usually show less difference; that station may only have a 50% greater incidence than other stations in the imputed database. The direction of the difference is right, but the degree is smaller; that’s attenuation, and it happens in most imputation applications.

## COMING NEXT

The next in this series of NPR White Papers will examine other ways to look at Arbitron data, including such measures as Listener Hours, Loyalty, Power, and Core Listeners. These measures have unique characteristics, and we’ll cover their usage and their caveats in White Paper #4.

# Radio Audience Estimates



**What They Are,  
Where They Come From,  
And How To Use  
(And Not Use) Them**

## CHAPTER 4: SUPPLEMENTAL MEASURES

Prepared for

National Public Radio

January 2005

*James D. Peacock  
Peacock Research, Inc.*

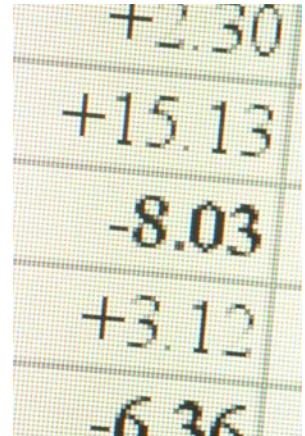
# TABLE OF CONTENTS

Chapter 4: Supplemental Measures .....	1
Table of Contents .....	2
Supplemental Measures .....	3
Listener Hours .....	3
Loyalty .....	4
Power.....	5
Core/"P1"/First Preference Listeners .....	6
Occasions Per Week .....	7
Exclusive Cume .....	7
Measures Used By Advertisers .....	8
Gross Rating Points/Gross Impressions .....	8
Reach and Frequency .....	8

In this chapter, we'll talk about a series of alternative measures that are often computed from Arbitron audience data by public radio researchers. These measures have unique applications and limitations, too, which are articulated below.

## SUPPLEMENTAL MEASURES

In addition to the main building blocks of AQH, Cume, and TSL (and the computational variations of ratings, shares, and persons), there are any number of other ways to look at Arbitron data. Of course, all these variations have their roots in the same basic diarykeeper data—entries of which station, and for how long. But other calculations have their place in the analytical toolkit.



### Listener Hours

One measure frequently used in public radio is called “Listener Hours.” This is a variation of the AQH Persons measure we already described, so most of the characteristics of AQH Persons data apply here too.<sup>1</sup>

Listener Hours are simply the number of AQH Persons (as reported by Arbitron) multiplied by the number of hours in the corresponding daypart. So, for example, if Arbitron reported that a station had 10,000 AQH Persons for Morning Drive (Monday-Friday 6-10AM), that would translate into 200,000 “Listener Hours” (10,000 persons times 20 hours).

Mathematically and otherwise, this measure has almost the same behavior as AQH Persons—except that it's bigger. The primary component of Listener Hours is the number of people estimated to be listening in any given quarter-hour. Therefore, like other AQH measures, increases in Listener Hours can be driven by reaching more people, or by getting them to listen longer, or both, and changes in Listener Hours need to be examined carefully to determine the underlying drivers.

But Listener Hours have one useful attribute: This measure can be used to present the relative audience value of two disparate dayparts. For example: Suppose Program A and Program B have identical AQH audiences. Now suppose that Program B airs for twice as many hours as Program A. It would be fair to say that Program B contributes more to the station's success than does the shorter Program A since it delivers the same average audience for a longer period of time. And that would show up in Program B having twice the Listener Hours of Program A.

The downside is that having twice the Listener Hours may or may not translate into having twice the true “value” to the station. Rather, it simply means that Program B delivers the same average audience for more hours during the week.

A related application is to compare a *program's* Listener Hours to a *station's* total Listener Hours. This, too, can be thought of as a “value” exercise—of all the “listening” done to the station, how much of it comes from that program? The numerical answer is a combination of audience success and the time devoted to the program.

Remember, though, that there are other ways to consider value and contribution. For example, a program could be responsible for bringing many new listeners to the station (i.e., the

---

<sup>1</sup> In some ways, this estimate is similar to a measure called Gross Impressions for ad campaigns, which is covered in a later section. The element in common is the simple addition of listening across all quarter-hours without regard for duplication.

program is building Cume), but if those new listeners listen for shorter periods of time, the Listener Hours measure may understate the audience-building contributions of that program.

Furthermore, a comparison of Listener Hours between two program elements doesn't immediately tell you whether the difference is driven by different average audiences, or simply by different durations on the air. As in the example above where a program has twice the Listener Hours simply by being on the air for twice as many hours, the Listener Hours number doesn't really tell you whether one program is performing better than another during the hours that each is on the air.

This measure seems to be unique to public radio, in addition to being well-entrenched.

However, users are encouraged to be cautious about using this hybrid measure too extensively. If the need is truly for a "value" measure—one that simultaneously considers both audience delivery and scheduled airtime—Listener Hours have utility. But for most other applications, the simpler measures of audience delivery would seem less likely to be misinterpreted.

## Loyalty

The measure called "Loyalty" is another audience estimate that's mostly limited to public radio, though all broadcasters tend to look at this issue of preference domination in some way.

The Loyalty measure is defined this way by Audience Research Analysis:

"Loyalty of the [station's] listener is the percentage of all his or her listening to radio that is to the supported station." [from Audience 98]

This measure shares many attributes with the standard Arbitron AQH share, but with one key difference. The Loyalty measure only considers in the denominator those who ever listen to the station in question (more precisely, a station's weekly Cume audience). Where an Arbitron share is on a base of all people listening to radio during the relevant time period, the Loyalty measure is on a base of all people listening to radio during the relevant time period *who also listen to the relevant public station at least once during the week*.

Here's an illustration of how these measures are related, using a market size of 1,000,000 population 12+, and about 2,000 diarykeepers over 12 weeks:

- 1) WAAA, a public radio station, reaches 100,000 different people (Cume) over the course of a week.
- 2) Those people reached in a week by WAAA represent 10% of the population of 1,000,000.
- 3) That weekly Cume audience of 100,000 also represents about 10% of diarykeepers, or 200 diaries.
- 4) During an Average Quarter Hour, about 20% of WAAA's weekly Cume audience is actually listening to the radio. That represents 20,000 people (20% of #1), and about 40 diarykeepers.
- 5) And during an Average Quarter Hour, about half of those 20,000 people listening to radio are actually listening to WAAA. That represents 10,000 people, and about 20 diarykeepers.
- 6) Loyalty is then computed as #5 divided by #4: 10,000 AQH WAAA listeners ÷ 20,000 WAAA Cume audience that are listening to *any* radio on an AQH basis = 50%. That measure is based on the 40 diaries that qualified in #4.

The Loyalty measure arose from the nature of public radio's access to Arbitron data in the past. Historically, public radio users only had access to detailed data about public radio listeners, and not to the entire Arbitron database. Thus, certain measures of "share" had to be computed on a base of only those public station listeners.

Loyalty has many pros and cons in common with Arbitron's standard AQH share, but with an important extra limitation: Loyalty is a less reliable estimate (i.e., it has a larger sampling error) than does an Arbitron share, which in turn is somewhat less reliable than an AQH rating. That's because the denominator of Loyalty is a relatively small number of diarykeepers—those who ever listen to the station *and* who are listening to the radio during an average quarter-hour of that daypart. (In the example above, Loyalty was computed on a base of 40 diarykeepers.)

Unfortunately, there's no easy way to estimate the amount of sampling error around a Loyalty score. But what's certain is that it's significantly less stable than a standard AQH share or rating.

That doesn't mean that Loyalty scores are without value. Some programming analysis will understandably focus on the listeners that a station already has, and that's what Loyalty does. It doesn't consider an expansion or contraction of the number of listeners (Cume), but it does tell you something about the allocation of listening among existing listeners.

In particular, it can be useful to trend a station's Loyalty estimates over time as an indicator of how much a station satisfies the radio needs of people who are at least occasional samplers of the station.

However, because of the problem with poor reliability, this author believes that Loyalty scores should always be accompanied by a statement of the number of diarykeepers in the denominator of the estimate. That way the users can decide for themselves if the base of the score is reasonable.<sup>2</sup>

In addition to reliability concerns, users of Loyalty scores also need to apply the same logic about expectations that would apply to standard AQH shares. For example, it's likely that Loyalty scores for a News/Talk station will differ by daypart, despite a station's best efforts. Even though the base of a Loyalty score consists only of a station's listeners, it still seems reasonable to expect that those listeners will vary in their preference for that type of programming by time of day.

Finally, and most importantly, Loyalty analysis should never occur in a vacuum or to the exclusion of measures of other behavior. Any "share" measure only tells part of the story of a station's progress. Loyalty scores are essentially AQH numbers, and as with other AQH estimates, there are several ways to achieve a particular Loyalty score; factors such as Cume and TSL should also be considered. For example, increases in Loyalty scores could mask decreases in total listeners reached (Cume); similarly, a conscious drive to increase the number of total listeners reached could correlate with a decrease in Loyalty, at least for some period of time.

## Power

Public radio literature also frequently refers to a measure called "Power." This statistic is a derivative of Loyalty, which means that it too is a variation on Arbitron's AQH data.

As defined by ARA, "A program's power is its ability to serve a station's listeners relative to all programming on the station. Technically speaking, power is defined as the weekly audience's loyalty to the program in relation to its loyalty to the station across the week."<sup>3</sup>

---

<sup>2</sup> Frankly, it should be standard practice to show the base—the number of diarykeepers—for any audience estimate presentation. But that knowledge is especially relevant to the small sample sizes involved with Loyalty scores.

<sup>3</sup> From the ARA publication, "Power and Affinity: Incorporating Two New Statistics Into Program Decision-Making."

Power is a relative measure, usually used to contrast a program or daypart's Loyalty score to that of a station overall. For example, if a station's Morning Drive programming has a Loyalty score of 40%, and the station's total-week, total-day Loyalty score is 30%, Power could be expressed as an index of 133 (40/30 times 100), or as a differential of +33%.

Conceptually, we could do much the same thing with standard Arbitron AQH data. Divide an Arbitron AQH share for a program with the station's overall AQH share, and we'd have something similar to the Power score, except that the base would be the total population.

In *Audience 98*, it was said that "a program's Power is its ability to draw listeners to the station. It is a measure of quantity, of strength." But we need to remember what lies behind the Power statistic; it's an index of two relatively unreliable AQH numbers (the two Loyalty scores) for two differently-defined dayparts. At a minimum, the same guidance applies to Power that applied to Loyalty: Power scores should always be accompanied by a statement of the number of diarykeepers in the numerator and in the denominator of the estimate.

The Power measure is also inherently affected by expectations, as discussed with Loyalty. The fact that an "off" daypart has a lower Power score than the station's average may be inevitable with that format.

### Core/"P1"/First Preference Listeners

The average Arbitron diarykeeper listens to about three stations per week. But some stations are chosen more often than others, and both commercial and public broadcasters are interested in learning more about listeners that choose their station as a favorite.

The most common way to segment listeners is by which station is listened to the most. A station's "Core Listeners" (or "P1/First Preference" listeners in Arbitron argot) are those who report listening to that station more than any other. This group is a subset of a station's total Cume audience; typically, these listeners represent from a third to a half of a station's total Cume listeners, but they can account for 70% or more of a station's Time Spent Listening (and of their AQH audience).

Public radio's "Core Listener" concept differs a bit from Arbitron's (and commercial radio's) "P1/First Preference" estimates. When Arbitron reports P1 data, the listeners are segmented uniquely for each daypart. A station's P1 listener is defined by that person's listening *in the defined daypart*; if the station is a respondent's primary station *in Morning Drive*, then the listener is a P1 listener *for Morning Drive*.

However, a public radio Core Listener is defined only once on a basis of their total-week listening (in Quarter Hours). If the listener spends a plurality of their total-week listening to that station, then that person is considered a Core Listener for that station in any analysis.

Here's an illustration:

Diarykeeper John Doe listens to 10 hours of radio during his diary-keeping week, distributed as:

WAAA	5 hours
WBBB	4 hours
WCCC	1 hour

John Doe would thus be counted as a Core Listener to station WAAA. (He would also count as a P1 listener in Arbitron for a full-week, total-day analysis.)

It appears that public radio listeners are somewhat more likely to be Core Listeners than are commercial radio listeners. The latest "Public Radio Tracking Study" reported that "The percentage of public radio listeners who are core to a station has risen dramatically, approaching 50 percent core composition." But one source of commercial station analysis (The Research Director, Inc.) indicates that 36% is typical of nonpublic stations on a total-day basis.

That means that people who sample public radio are more likely to have a public station as their favorite.

It's been established in other studies that public radio Core Listeners are also the most likely to be financial contributors. Therefore, it makes sense to track the number of Core Listeners over time, and to better understand their characteristics. In fact, the Fall 2003 "Public Radio Tracking Study" said, "If we had to pick a single number to assess and trend the performance of public radio stations, that metric would be the number of Core listeners."

Core Listener analysis has a few caveats. For one, remember that fewer diarykeepers are involved; the number of actual diarykeepers in a market that choose one public station as their favorite is often a relatively small number, so earlier cautions about sample sizes apply here too. Presentation of the actual number of tabulated diarykeepers should be a requirement for such analysis.

Analysts should also remember that a Core Listener isn't necessarily a *heavy* listener to the station. Even light radio listeners are Core to some station.

And under the heading of expectations, it's also likely that different formats will have different Core Listener compositions. What's normal or achievable for a Classical station may well be different than for a News station; format norms should be the benchmark, not overall averages.

Finally, as with the relationship between Cume and TSL, remember that acquiring new listeners (i.e., building Cume) can cause at least a temporary reduction in Core Listener percentages (though hopefully, not in the *number* of Core Listeners).

## Occasions Per Week

We mentioned a moment ago that the typical Arbitron diarykeeper mentions about three different stations over the course of a week. It's also true that they listen to radio about twice a day, although that can vary widely by demographic and other characteristics. (Arbitron reports that the average diary contains 15 entries.<sup>4</sup>)

It's sometimes useful to analyze these "occasions per week," thinking of each discrete diary entry as an occasion of radio listening. While this number isn't a "standard" Arbitron estimate, it can help programmers to remember how listeners actually use the radio, and their stations. Most listeners don't sit glued to their radio all the time; they come and go as their lives and needs warrant.

Because this number has intuitive meaning, it can be a helpful way to think about listener behavior (or at least, about Arbitron crediting of listener behavior). Just remember that reliability can be an issue with this number since the "base" is "number of listeners," not the total number of people surveyed.

And like many other audience estimates, the "norms" for Listening Occasions can vary significantly by demographic, from format to format, and across dayparts.

## Exclusive Cume

Arbitron tabulates an estimate called Exclusive Cume for a variety of dayparts. This number is reasonably self-explanatory—it's the estimated number of people who listen *exclusively* to a

---

<sup>4</sup> Remember that Arbitron gives no credit for listening occasions of less than five minutes duration, and that a single quarter-hour credit can occasionally represent more than one entry.

particular station during a particular time period. It tends to be most useful when examining specific dayparts; the number of people who are exclusive to a station over an entire week is relatively small.

This number is a loose correlate of the Core Listener estimate discussed earlier. In a sense, we could almost consider these folks Core Cumers. But while Core Listener estimates are based on the amount of time spent with a station, Exclusive Cume estimates (like other Cumes) only consider whether or not someone listened to a station regardless of duration.

There's always some debate about the value of an exclusive listener. Since we don't know whether an Exclusive Cume diarykeeper listened a little or a lot, their real value to the station is murky.

Similarly, it's debatable whether a station should strive for a large number or percentage of Exclusive Cume listeners. Does a high percentage of Exclusive listeners indicate above-average satisfaction, or does it simply reflect programming with a very narrow appeal?

In and of itself, Exclusive Cume doesn't provide much guidance.

## Measures Used By Advertisers

For the record, let's also document a few other measures mostly used by advertisers in planning and buying radio (and other media).

### GROSS RATING POINTS/GROSS IMPRESSIONS

We noted in an earlier chapter that AQH Ratings and AQH Persons are used by advertisers to estimate the number of people who were probably exposed to a single commercial. But of course, advertisers rarely run just one ad; the real interest is in the number of people exposed to an entire campaign of multiple ads.

One way to "size" a campaign's delivery is to simply add up the audiences for each commercial that ran. That's the logic behind Gross Rating Points (GRPs) and Gross Impressions. The former number simply adds up all the AQH Ratings for each spot; the latter adds up the Persons Estimates for each individual ad.

The term "gross" is well chosen, since this is a very crude and significantly overstated number. Adding up all the individual ratings or projections fails to account for duplication—for the fact that the same person may be exposed to an ad multiple times. If Spot #1 reached 10,000 people and Spot #2 also reached 10,000 people, the Gross Impressions would equal 20,000. But we can be reasonably confident that the second group of 10,000 people overlaps with the first; the *net* number of people reached by both ads would be something less than 20,000.

Obviously, then, GRPs and Gross Impressions are rather clumsy ways to consider the actual ad impressions of a campaign. But the measures persist, mostly because they are easy to compute; accounting for duplication requires more than just the published data.

Also note that GRPs and Gross Impressions have some unique statistical attributes. If you need to estimate the reliability of those numbers, be sure to read Arbitron's publication called "Arbitron Study of Radio-Schedule Audience Estimate Reliability," available at Arbitron's website.

### REACH AND FREQUENCY

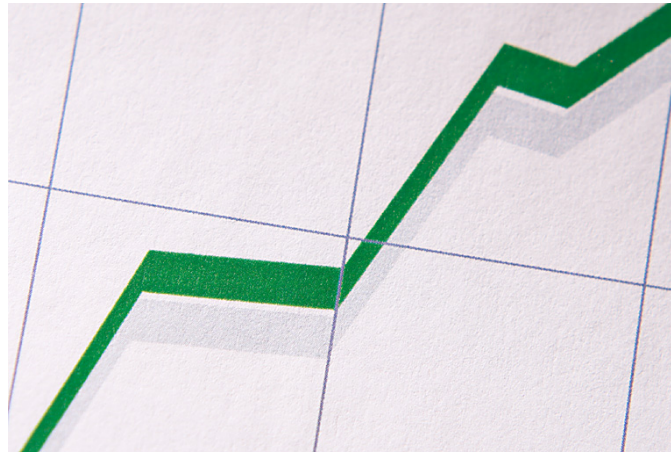
Because GRPs and Gross Impressions are crude and overstated, methods and tools have arisen to estimate commercial campaign audiences more precisely. These tools provide estimates called Reach and Frequency—the number of different people reached by a multi-spot campaign (Reach), and the number of times each of those people was likely to have heard the spot (Frequency).

Mathematical models are usually used to make such estimates because of reliability problems with Arbitron data. Remember that Arbitron usually reports 12-week averages for station ratings; those ratings are the average of 12 discrete one-week surveys. To compute the actual reach of a single ad on a single day would require using the diaries from only one week—not a very stable estimate. Though such numbers can be computed, practitioners usually prefer the simplicity and stability of model-based estimates which derive Reach and Frequency estimates from published 12-week averages.

Fundamentally, of course, we've seen one of these numbers before; Reach is the same concept as Cume, except that we're estimating Cume for a collection of specific moments in time. Frequency is a derivative of Time Spent Listening.

These numbers also have unique statistical properties, also documented in the Arbitron document mentioned above.

# Radio Audience Estimates



**What They Are,  
Where They Come From,  
And How To Use  
(And Not Use) Them**

## CHAPTER 5: INTRODUCTION TO RELIABILITY

Prepared for

National Public Radio

January 2005

*James D. Peacock  
Peacock Research, Inc.*

# TABLE OF CONTENTS

Chapter 5: Introduction to Reliability .....	1
Table of Contents .....	2
Introduction To Survey Reliability .....	3
What's A "Standard Error," And Why Do We Care? .....	3
Factors Which Affect Reliability .....	4
Some Practical Advice On Reliability .....	6

In this chapter, we'll be a little more specific about the reliability issues that affect radio audience estimates. For this paper, the word "reliability" is used to denote relative stability, or what's often referred to as sampling error.

## INTRODUCTION TO SURVEY RELIABILITY

In the preceding papers we referred to "reliability issues" several times. Up until now, we've mostly just talked about reliability as being synonymous with stability. Surveys are based on samples of the populations, and results from surveys can "bounce around" simply because we haven't surveyed the entire population.

### What's A "Standard Error," And Why Do We Care?

Before we leap into the complexities of estimating reliability, we should spend a moment on terminology. The reliability of survey-based estimates can be quantified through something called the "Standard Error," which allows the construction of "confidence intervals." This section will try to explain both terms.



The Standard Error has little intuitive meaning, unfortunately. It's simply a particular mathematical way of expressing the stability of estimates from samples of particular sizes. But one aspect of Standard Errors is easy to remember—a rating with a larger Standard Error is less reliable (i.e., is more likely to be bouncy) than a rating with a smaller Standard Error. All else being equal, the smaller the Standard Error, the more confidence you have in the stability of ratings from a particular sample. [We'll talk more in a moment about how Standard Errors are calculated.]

Note that the Standard Error only quantifies the stability of a rating. It is not an estimate of the other kinds of error that can affect radio estimates. Only sampling error is measured by a Standard Error; other kinds of "error" aren't taken into account.

So aside from knowing that "smaller is better," what's the utility of a Standard Error for a particular rating? That's where "confidence intervals" come in. A confidence interval is a "plus or minus" range that describes the potential bounce around a rating. For example, a "68% confidence interval" is that range of values which is 68% likely to include the rating you'd get from a survey of the total population—or "truth," as well as this measurement method could measure truth.

Suppose, for example, you were told that a rating of 2.0 from the current Arbitron sample had a 68% confidence interval of 0.5. That would mean that there's a 68% chance that the total population's rating would be between 1.5 and 2.5, had Arbitron measured the whole population instead of just a sample.

Another way of putting it is that if Arbitron had conducted an infinite number of identical surveys at the same time, but with different samples, 68% of them would have produced ratings between 1.5 and 2.5. Of course, a few of those surveys (32%) would have fallen outside that range.

Fortunately, Standard Errors are fairly easy to calculate for the most common Arbitron estimates, as you'll see in a moment. And even more fortunately, they have a predictable relationship with confidence intervals. Specifically:

A confidence interval of this percentage:

Equals this number of Standard Errors:

68% confidence	=	1.00	Standard Errors
80% confidence	=	1.28	Standard Errors
90% confidence	=	1.65	Standard Errors
95% confidence	=	1.96	Standard Errors
95.5% confidence	=	2.00	Standard Errors
99% confidence	=	2.58	Standard Errors

As you can see, the higher the confidence level, the wider the interval. You can be 99% confident in a fairly wide range; a narrower numerical range is accompanied by a lower degree of confidence.

To return to our example of a moment ago: If the 68% confidence interval of a 2.0 rating was 0.5 points, then the 95.5% confidence interval would be exactly twice as large, or 1.0 points. That's because the 95.5% confidence interval equals 2.00 Standard Errors, or 2.0%  $\pm 0.5$ .

Of course, it's possible that the rating from the sample exactly matches the total population's "true" rating. But we can never be sure since we're always looking at ratings from samples; truth is regrettably elusive.

In short: Ratings from a sample (like those of Arbitron) are always estimates of the population's behavior. The population's real behavior is probably close to the sample's if the sample is designed well; in fact, you can be 68% sure that the population's real behavior is within one Standard Error of the rating from the survey. But the survey data are still estimates—estimates whose reliability can be described through Standard Errors and confidence intervals.

## Factors Which Affect Reliability

Arbitron surveys are designed to be probability samples which are projectable to the defined universe. A probability sample is a prerequisite for developing realistic estimates of the amount of sampling error, or "bounce," associated with that data.

The size of the sample upon which a rating is based is a key part of the rating's stability; larger samples yield more reliable ratings. But many factors beyond sample size influence the true reliability of a real-world sample. Among them are:

- **Sample Stratification.** Samples which are stratified (controlled) to demographic or geographical targets can be more reliable than samples which are "simple random samples" (SRS), since some of the potential data variance is controlled. Assuming that the control variables are associated with the behavior being measured, such control makes it less likely that the sample will differ wildly from the population or from future samples. (Arbitron stratifies its samples by county and related geographies.)
- **Sample Clustering.** Sample designs which select multiple persons per household can lose some reliability for demographic measures because of the household clustering

of the sampled persons, since there's a correlation of behavior among household residents. (Arbitron measures multiple people per household.)

- **Weighting or Sample Balancing.** Real-world samples are often weighted in the process of computing the reported estimates, so that underrepresented groups are weighted up and overrepresented groups are weighted down. Weighting can reduce the potential bias from under- and overrepresentation of demographic groups, again assuming that the variables used are correlated with the behavior being measured. But weighting can also reduce the reliability of a sample by increasing the potential impact of atypical households or persons.
- **Sample Homogeneity.** In media research, it's common to examine data for narrowly defined demographic groups. It's often the case that data from narrower, demographically-similar groups will be relatively more reliable than data from the population at large, all else being equal. For example, Women 18-34 listen to a smaller array of stations and programs than the population at large; that reduces the potential variance in audience estimates for that group compared to a broader demographic.
- **Repeated Measures.** Many common audience measures, especially those which represent averages of time-spent estimates (e.g., Average Quarter Hour), can benefit from the averaging process. An AQH rating for a full week, for example, actually represents an average of many different observations of each person's behavior. Thus, average-point-in-time measures can often be significantly more reliable than Cume measures, and averages for long time periods are more reliable than averages for short periods.
- **Rating Size.** Of course, the absolute size of an audience estimate is also relevant to its reliability. All else being equal, a larger rating will have a relatively smaller Standard Error than will a smaller rating from the same sample. While the absolute Standard Errors increase as the rating size increases, the Standard Error represents a declining proportion of the rating. For example, if a 1.0 rating from a sample of 1000 has a Standard Error of 0.3 (30%), a 10.0 rating from that same sample would have a Standard Error of 0.9, or only 9%.

All of these factors combine to make real-world samples behave differently than most textbook formulas would predict. Those differences are often summarized in the measure called "Statistical Efficiency," which describes whether a particular estimate has a real Standard Error that's larger or smaller than simple-random-sample (SRS) theory would estimate. A Statistical Efficiency of 1.25 indicates that the rating in question has reliability equal to an SRS sample that's 25% larger. Statistical Efficiencies of less than 1.0 indicate that a particular sample is less reliable than an SRS sample of the same size.

The true reliability of a particular method must actually be estimated empirically through such techniques as jackknife replication. That's the method used by Arbitron to provide the "Table A" and "Table B" values which accompany each market report, and users are encouraged to become more familiar with Arbitron's recommended procedures for estimating actual Standard Errors.

The study on which Arbitron's reliability model was based is regrettably old, but it's still a reasonable approach to estimating the reliability of radio audience estimates. Like the audience estimates themselves, Standard Errors are estimates, too, and should not be considered to have a great deal of precision.

## Some Practical Advice On Reliability

In addition to the technical details above, here's a more practical summary of how reliability issues affect radio analysis. In a nutshell:

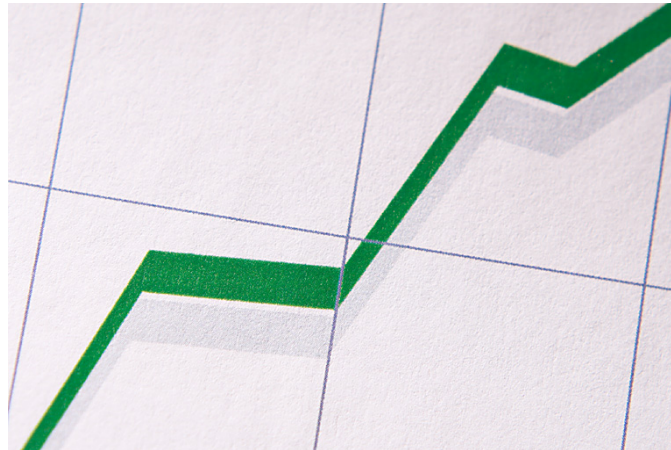
- The important thing to remember is that every number derived from an Arbitron survey has sampling error. If nothing else, make sure that the number of diaries forming the basis for the tabulation is prominently available and considered.
- Remember that the diary count which matters is the count of diaries in the denominator (the "base") of the calculation. If the audience estimate is an AQH rating, the appropriate base is all diarykeepers in that demographic. If the estimate is an AQH Share, the denominator is smaller—the number of diarykeepers *who were listening in that daypart on an AQH basis*. If the estimate is Loyalty, the base is smaller still—the number of diarykeepers listening in that daypart on an AQH basis *who also listened to that station at all during the week*.
- All else being equal, AQH estimates are more reliable than Cume estimates of the same behavior. (That's because of the repeated-measures phenomenon mentioned above.)
- Smaller demographics have less reliable estimates than broader demographics because of the reduced sample size, but the loss of reliability is partially mitigated by other factors (homogeneity in particular).
- All else being equal, broader dayparts (like Mon-Sun 6AM-Midnight) yield more reliable audience estimates than briefer dayparts (e.g., Morning Drive). Single-hour estimates are the least reliable of all.

More than anything, don't assume that sample reliability is irrelevant. When all else fails, make sure you know how many diaries were involved, and make an informed judgment about the linked data.

Finally: This section should help the reader understand why there's no simple answer to the question, "How many diaries are 'enough' to be reliable?" While there are some common conventions in the industry (like requiring a minimum of 30 diaries to be "usable"), the reality is that there's no one simple dividing line between reliable and unreliable audience estimates. Because of all the factors above, for example, a 12+ Monday-Sunday 6AM-Midnight number based on 30 diaries would have a different standard error than would a similar 30-diary estimate for a narrower daypart or for a narrow demographic.

In the end, common sense should prevail—*informed* common sense, based on full disclosure of the amount of tabulated sample. If the sample seems too small for an important decision, it probably is.

# Radio Audience Estimates



**What They Are,  
Where They Come From,  
And How To Use  
(And Not Use) Them**

## CHAPTER 6: OTHER RADIO DATA SOURCES

Prepared for

National Public Radio

January 2005

*James D. Peacock  
Peacock Research, Inc.*

# TABLE OF CONTENTS

Chapter 6: Other Radio Data Sources .....	1
Table of Contents .....	2
Other Syndicated Surveys With Radio Data .....	3
MRI .....	3
Scarborough .....	4
The Media Audit (International Demographics) .....	5
Simmons (SMRB) .....	5
Electronic Radio Measurement .....	6
PPM And What It Means For Radio .....	6
Navigauge .....	7
Final Summary .....	8
About The Author .....	9

Arbitron is not alone in measuring radio listening. In this final chapter of six on radio audience estimates, we'll touch briefly on a few other syndicated research companies that have radio data available.

## OTHER SYNDICATED SURVEYS WITH RADIO DATA

### MRI

MRI (Mediamark Research, Inc.) is the primary provider of magazine audience measurement in media research. But because MRI also collects a great deal of other data, they are often used for multi-media planning and for qualitative profiling.

**Mediamark Research Inc.**

Overall, MRI is a very high quality source of data. Like Arbitron, they're Accredited by the Media Rating Council, the industry's rating-service watchdog. MRI also has the highest response rate of any syndicated media research, even for their detailed product and qualitative data.

Despite their overall quality, however, MRI doesn't pretend to be the definitive source of radio data. For starters, MRI is primarily a nationwide measurement service oriented toward national magazines. Only limited data are available at the local level, and only in the largest markets.

It's also well known that multi-media measurement services like MRI will never yield audience estimates which are identical to those of a single-medium service like Arbitron.<sup>1</sup> Research has shown that the very process of measuring multiple media can affect the data reported about any one medium.

MRI is also much more limited in its ability to interpret ambiguous entries than is Arbitron. Because MRI is not a local-market service, it doesn't maintain as much information about each station in each market; as a result MRI has more uncredited listening than Arbitron.

Finally, the MRI method for collecting radio data is quite different from Arbitron's. As described by MRI, which collects the radio data during an in-person interview:

*The interviewer displays cards on which are listed five time periods. While showing this card, the following questions are asked:*

*"These are time periods during which people can listen to or hear a radio. To the nearest half hour, how much time, if any did you spend listening to or hearing a radio in each of these time periods yesterday, either in your home, car or any other place?" and "During the period (time period), what station or stations did you listen to? Please give me the call letters of each station and whether it is AM or FM." These two questions are asked for "yesterday," for "last Saturday" and for "last Sunday."*

Without getting into all the pros and cons of MRI's method compared to Arbitron's, it's guaranteed that the MRI method will yield different radio audience estimates. It should not be considered a definitive source of radio data, though its ability to provide more characteristics about radio listeners is frequently valuable.

---

<sup>1</sup> MRI's magazine data is collected first in their method and is presumably untainted by the burdens of the following questions. But MRI data about other media are collected among many other questions.

## Scarborough



Like MRI, Scarborough research services are known in part for offering a wealth of data about multiple media and about consumer behavior. While Scarborough has its origins as a newspaper measurement service, it has grown into an incredibly complex product that collects a wide array of data. Some data are collected over the telephone; other types are collected via a very long written questionnaire; and most TV data are collected by a written diary.

In part because of that complexity, Scarborough data are only partially Accredited by the MRC as of this writing. Data collected by telephone—the newspaper and basic radio data—are presently Accredited, as are the data collected from the long written questionnaire; data based on the TV diary are not yet Accredited, though Scarborough continues to refine those procedures in hopes of achieving Accreditation.

The radio audience measures collected by Scarborough use a very different method than Arbitron's. Modeled on the telephone-recall methods used by the now-defunct Birch ratings service, Scarborough uses a telephone call to ask people about their radio listening for "yesterday." Questions are also asked about any radio stations sampled during the last week.

Without going into all the details, Scarborough collects data that allow computation of AQH estimates, but only based on one day's worth of listening from each respondent. They also collect limited Cume data for weekdays, weekends, and total week.

Scarborough has access to Arbitron's database of information about station characteristics, which improves Scarborough's ability to credit any ambiguous entries. Interviewers also probe any mentions which aren't clear.

However, a unique characteristic of Scarborough's method is an attempt to "conform" their radio estimates to Arbitron's. Radio audience estimates based on a telephone recall method like Scarborough's will inevitably differ from those of Arbitron's diary method. Practitioners sometimes argue about the pros and cons of each method, but the bottom line is that Scarborough's radio data (before processing) differs from Arbitron in scope, amount, and distribution.

Since Scarborough is half-owned by Arbitron,<sup>2</sup> and since the marketplace would rather not have two sets of differing radio ratings to contend with, Scarborough "conforms" their raw data to almost perfectly match Arbitron's for most common measures. Therefore, Scarborough's basic radio data (as published) offer no advantage over Arbitron's.

However, Scarborough data still have appeal for radio because of the potential to cross-tab the radio data with the extensive data collected elsewhere in the Scarborough survey. Almost any type of consumer behavior, and many kinds of behavior with other media, can be linked back to radio listening in the Scarborough database.

While that capability has great appeal (and fairly wide use in the commercial radio market), strict researchers have serious reservations about such applications. Scarborough's written questionnaire and TV diary have very low response rates; furthermore, Scarborough uses very elaborate imputation procedures to fit all the types of data together.

This is a classic trade-off in survey research. The more data you try to collect from one respondent, the more suspect the final results are because of reduced cooperation. Increased database utility comes at the expense of sample representativeness.

Because Scarborough offers many types of cross-tabs not available anywhere else, many users accept the trade-off. But wise researchers should always remember that the data represent a relatively small percent of the population—those who are willing to answer that many questions.

---

<sup>2</sup> Nielsen's owner, VNU, owns the other half of Scarborough.

One side note concerning the Scarborough service: Because of its newspaper origins, Scarborough is primarily a local-market service. Its most commonly used products are based on local-market surveys in the top 75 television markets. Those markets are defined as larger geographies than the Arbitron Metros, referred to as Designated Market Areas, or DMAs.

Scarborough has recently added limited additional sample outside those 75 markets in order to offer national data based on sample across the country ("Scarborough USA+"). This product has not yet been submitted for Accreditation by the MRC. In using this product, remember that the counties outside the top 75 markets are measured with a relatively small amount of sample.

## The Media Audit (International Demographics)

The Media Audit (TMA) is a product of the research company International Demographics, Inc. TMA received renewed attention in the radio industry when Infinity Broadcasting recently threatened not to renew its Arbitron contract in favor of using TMA in many of its markets.



Unlike Scarborough, TMA is a relatively simple product. Though TMA does collect data about multiple media and about consumer behavior, all measures are collected in a single telephone interview of about half an hour in duration. TMA surveys are usually restricted to Metro-defined geographies, and TMA is now present in about 80 markets. Each market can be measured either once or twice per year, each time for a three or four week survey period.

TMA collects less specific radio data than Arbitron. Telephone participants are asked to recall all stations that they listened to for five minutes or more over the last week (a weekly Cume measure). They're also asked to identify which of those stations was listened to the most. In addition, the amount of listening to any radio is collected for each time of day "yesterday"; this yesterday-listening section is not station-specific.

The measures for other media are similarly less detailed than are the measures usually used Nielsen, Scarborough, MRI, etc., though local broadcasters do benefit from TMA's collection of section-specific newspaper data. (Most newspaper data from other sources measures exposure to an entire edition of a newspaper, not to specific sections.)

Methodologically, TMA's greatest weakness is its low response rates. That's partly a function of the length of the questionnaire. It also reflects some economizing on procedures that might improve response.

The Media Audit is no longer Accredited by the MRC, after voluntarily withdrawing from the Accreditation process.

## Simmons (SMRB)

The nature of the Simmons product has changed significantly over the years as ownership and management of the company have changed. The primary Simmons product is now the Simmons Unified National Consumer Study, based on a telephone-placed self-administered questionnaire. The operational procedures are similar to Arbitron's—a telephone call to solicit cooperation, followed by a mailed-out, mailed-back questionnaire.



But the Simmons questionnaire is quite long, measuring 8,000 brands in 450 product categories. In addition, Simmons includes over 600 lifestyle and opinion questions, and the media data include "virtually every network and cable television show, more than 40 radio formats, over 200 magazine titles, all major national newspapers and the top 75 Internet sites" (quoted from the Simmons website).

Simmons data are collected year-round, but are reported only for six- and 12-month periods, with two-year averages available for narrow categories. Simmons also provides special

surveys of teens and children, and supplements its national sample with additional Hispanics to permit more analysis of that population.

Of key interest to radio users is that the radio data are limited to formats at the national level. Users can access local-market subsets of the national data, but the radio data are still only format-level.

Simmons is actively seeking customers for a reinterview-survey product, in which the existing Simmons samples can be queried a second time to collect additional, more specific data. But since low response rates are already a problem for the core Simmons product, a reinterview study is likely to suffer from serious problems of double-nonresponse.

## ELECTRONIC RADIO MEASUREMENT

### PPM And What It Means For Radio

Looking into the future, many broadcasters are hoping for an electronic alternative to Arbitron's paper diary methodology. Among other things, it's believed that radio audience measurement isn't taken as seriously as ratings for other media because of a perception that diaries are antiquated.



Arbitron itself is hoping to provide that alternative through its Portable People Meter, or PPM. In development since the early 1990's, the PPM is a pager-sized device which respondents would carry for an extended period of time; during that survey period, the PPM would detect in-audible audio codes embedded in the signals of participating radio stations (along with any other participating media that have encoded audio signals).

Arbitron has conducted several in-market tests of the PPM technique, most recently in conjunction with Nielsen Media Research. Arbitron believes that PPM can be a viable method for measuring radio in the U.S. if (and many would say, *only* if) the technique is simultaneously used for TV measurement. That's where Nielsen comes in; since Arbitron does not wish to compete with Nielsen for the TV ratings business, it has sold a "right of first refusal" to Nielsen for the rights to the TV data in the U.S.

The future of the PPM is unclear as of this writing. Though there seems to be interest in the methodology outside the U.S., domestic implementation faces two significant hurdles.

Though many people believe PPM could be more accurate than diaries, testing has shown that PPM is likely to change published ratings in significant ways. Station rank orders will change, and the relationship of TV and radio to each other could change. Since any such change would create winners and losers, Arbitron (and Nielsen) would have a major uphill battle to convince a majority of users that all the changes are reasonable and beneficial—especially if the cost of measurement goes up, as it almost certainly would with PPM.

Secondly: In the U.S., commercial deployment of PPM for radio seems completely dependent on simultaneous deployment for television, for cost reasons. And that means getting support from Nielsen. It's this author's opinion that Nielsen is unlikely to ever support commercial deployment of PPM, which would likely doom PPM for radio—unless Arbitron finds some other way to get back into the TV ratings business.

Nevertheless, the progress of PPM is worth following. It does offer the exciting potential of some multimedia data from the same panel (at least TV and radio); it has some support among

major advertisers and agencies as a step forward for the credibility of radio; and in principle, it could be shown to be more accurate.

Aside from the business issues, researchers who scrutinize PPM are encouraged to study and understand Arbitron's latest developments in the following areas:

- Device sensitivity. There are open questions about whether the audio being recorded by the device is really similar to the audio being "heard" by respondents.
- Editing rules and compliance. Despite the "passive" nature of the data collection, there are still many areas where Arbitron must impose judgment in the form of editing rules. This is especially true when it comes to assessing respondent compliance, and the rules for deciding which respondents participated "well enough" can have large impact on the reported audience estimates.
- Response rates. Arbitron has reported significant progress in getting better respondent cooperation with PPM surveys. Those efforts will need closer scrutiny, especially as they relate to cooperation within specific population groups (ethnics, young adults and teens, etc.).

## Navigauge

Another company pursuing electronic measurement for radio goes by the name of Navigauge. This firm provides a service that passively measures in-car radio tuning, along with the geographical location of sampled cars through the use of GPS (global positioning satellites). To participants, this process is referred to as the IQMedia Panel.

As described on the Navigauge website:

*"Our innovative monitoring system is built around a patent-pending passive device that does not require any internal modifications to the vehicle's audio components. When connected to the vehicle audio system it automatically collects data on radio usage and vehicle location. Each change in the radio dial and vehicle position is accurately time-stamped and stored; the data is then transmitted over a national wireless communication network to the central servers in Navigauge's Network Operations Center. The data is securely stored and then processed by the Navigauge AARMS™ (Automated Audience Report Management System) for use by the client."*

Though the Navigauge system is appealingly passive, it is limited to in-car measurement. It also does not provide direct identification of the specific person driving the car at the time, which prevents positive association of tuning and demographics. The marketplace value of such data is not yet clear.

However, Navigauge has associated itself with a group of knowledgeable industry professionals, so they are definitely worth watching. It seems unlikely that Navigauge would supplant Arbitron as a source of overall listening data, but it could become a useful adjunct.

## FINAL SUMMARY

In this report, we set out to document some of the most common radio audience estimates used in public broadcasting, along with their uses and limits. We hope that users have gained a better feel for the myriad of numbers that can be derived from those deceptively simple entries made by survey respondents.

It would be difficult to summarize all the different recommendations made in this series, but some of the more important conclusions include these:

- Radio audience data are imperfect estimates of the past, not perfect predictors of the future.
- Arbitron's diary-based methodology remains the standard for sizing radio audiences, if not for describing all their characteristics.
- Despite all the many ways that we compute audience estimates, remember that the information provided by Arbitron diarykeepers is extremely basic—what station was heard, when, and where. The actual meaning of those behavioral measures is in the eye of the analyst.
- While radio audience estimates can be calculated with seemingly great precision, every audience estimate has significant sampling error around it. At a minimum, all radio estimate presentations should be accompanied by information about the amount of sample used in their calculation.
- We also need to remember that Arbitron's published data reflect an operational crediting process, which turns the exact times provided by diarykeepers into quarter-hour chunks of credit. Credits that appear continuous may not have been so in reality.
- Arbitron's primary audience numbers—the ubiquitous Average Quarter Hour estimates—have many virtues beyond those applicable to advertisers. AQH numbers are often the most statistically reliable, and their meaning is clear.
- But searches for causality do need to go beyond AQH. Audience size consists of both people and the amount of their listening, and strategic analysis should consider both individually. Such analysis also needs to remember that people reached and time-spent-listening are often inversely related over time—another reason why AQH is such an important benchmark.
- Our hunger for additional data comes at a cost. For example, it's possible to collect or impute more data about Arbitron respondents, but such secondary data collection brings additional important caveats.
- All else being equal, simple measures are better than those which are compound measures of distinct phenomena.
- And finally, analyst expectations are key. Before comparing two numbers to each other, we need to have reasonable expectations about what that relationship *should* be. Assumptions that appeal can be universal are dangerous.

## ABOUT THE AUTHOR

This document was prepared for National Public Radio by James Peacock, President of Peacock Research, Inc.

The company was founded in 1996, and it provides a full range of management and technical consulting to purchasers and managers of survey and media research. Mr. Peacock serves as research consultant to the Media Rating Council (MRC), the Radio Ad Effectiveness Lab (RAEL), and The Weather Channel, in addition to National Public Radio.

Prior to forming Peacock Research, he was VP/Research at Arbitron, leading the departments that were responsible for all methodology research including the development of survey and panel methods for new measurement technologies. He and his staff were responsible for day-to-day responses to customer inquiries on methodological issues, and for relations with industry research organizations.

In addition, he is the co-inventor of an Arbitron patent covering several approaches to portable electronic survey methods.

Mr. Peacock also spent seven years at Susquehanna Broadcasting Company, providing ratings analysis and developing primary research methods for music evaluation for the Radio Division.

Mr. Peacock holds an MBA from Southeastern University with a major in Marketing. He has written and spoken extensively for media research industry conferences including those sponsored by the RAB, the NAB, The Advertising Research Foundation, BBM Canada, IBOPE Latin America, ESOMAR, the Media Research Council of Chicago, and others.