



Radio-Schedule Audience Estimate Reliability

Final Report (September 1995)

ARBITRON

Arbitron Study of Radio-Schedule Audience Estimate Reliability

Final Report (September 1995)

Table of Contents

Introduction	1
Background	1
Measures	1
The Issues	2
Overall Summary of Results	3
Section One: Sampling Error Within One Survey	5
Study Scale	5
Reliability Estimation: A Few Technical Definitions	6
Variables Affecting GRP Reliability	7
Variables Affecting Net Reach Reliability	11
Variables Affecting Frequency Reliability	14
Computing Single-Survey Confidence Intervals	17
Estimating Single-Survey GRP Reliability (Confidence Intervals)	17
Estimating Single-Survey Reach Reliability (Confidence Intervals)	20
Estimating Single-Survey Frequency Reliability (Confidence Intervals)	22
Table 1.0: Statistical Efficiency for Single-Station Gross Rating Points	24
Table 1.1: Multiple-Station Adjustment Factor (MS) For GRPs	25
Table 2.0: Statistical Efficiency for Net Reach	26
Section Two: Sampling Error When Comparing Two Surveys	27
The Two-Survey Formula	28
Section Three: Variance from Comparing Hour-Based GRPs with Daypart Expectations ..	30
Some Examples	30
How to Compute This Effect for Other Situations	34



Section Four: Variance from Rounding in GRP Computation	36
Examples	38
The Formula	39
Examples of 95% Confidence Intervals for Rounding-Error Variance of Rating-Point-Based GRPs	41
Section Five: Technical Notes, Caveats and Other Miscellany	42
Averaging Books for More Reliability	42
Two-Tail vs. One-Tail Probabilities	42
Market-Level Variation in Efficiency	42
Reach vs. Cume	43
List of Markets	43
Applicability to Markets Ranked 50+	43
Acknowledgments	44
Appendix A: Adjustment Factors for Odd-Spot Variation	45
Appendix B: Restrictions and Other Limitations	48
General Information and Limitations	48
Disclaimer of Warranties	48
Limitation on Liability	49
Restrictions on Use of Report	49

Introduction

Background

During the early months of 1995, radio schedule audience estimate evaluation became a topic of considerable debate in the advertising industry. Some advertising agencies, on behalf of their advertiser clients, began showing increased interest in accountability measures for the delivery of audiences for local spot campaigns. Broadcasters, in turn, expressed a wide variety of concerns about how accountability analysis might be conducted and utilized, and about the appropriateness of that exercise for the medium (or at least, for the medium as it's approached today in the buy-sell process).

Amid the many controversies was one clear agreement: Both sides in the debate expressed a desire to know more about the reliability of Arbitron radio audience estimates when used in a schedule evaluation context. The American Association of Advertising Agencies, the Arbitron Radio Advisory Council, and several specific advertising agencies all requested a new study of the reliability of aggregated estimates of radio-schedule audience delivery. While it's clear that Arbitron has no role in recommending business-practice guidelines, Arbitron committed to a statistical study of reliability, and to share the results with the industry at large.

The initial specifications were developed with feedback from a wide range of industry researchers on how the study should be conducted, and additional significant input was received after we issued an Executive Summary of our preliminary findings in July 1995. That input helped us identify a number of complex issues that are described further below.

The primary study (of sampling error within one survey) was conducted by Dr. Roland Soong of Audits & Surveys Worldwide, under contract with Arbitron. Dr. Soong is an expert in the reliability of audience estimates, with several publications in this area. Additional analyses were conducted by Arbitron's Research Group, with input from Dr. Soong. The current report is our final planned output from this project.

Measures

Because the study attempted to meet the needs and interests of a variety of agencies and broadcasters, we considered the reliability of both Gross Rating Points (GRPs) as a measure of total delivery, and of Reach and Frequency as alternative measures.

In most of our analyses, the "delivery" of a schedule was estimated as follows. The GRP method attributed to each spot the Metro-level 12-week-average Arbitron Average Quarter Hour (AQH)

audience estimate for the daypart in which the spot ran; in Section Three, we discuss issues concerning the use of hour-by-hour data. Those daypart estimates were then summed across spots. (See Section Four for additional discussion of GRP computation methods and rounding error.)

The Reach and Frequency estimates were single-week estimates based on the partially-modeled algorithms used in Arbitron's Maximi\$er product as of Winter/Spring 1995. This Maximi\$er method uses a hypergeometric estimation model wherein each respondent is given a probability of exposure for the number of spots that ran within a standard daypart.

The Issues

As we began studying this subject, it became clear that there are several distinct components to the variance around radio schedule audience estimates. Any user of this study should consider each of these issues and its potential relevance to that user's practices:

- ▶ **Sampling Error within One Survey:** The first issue we studied in depth concerns the sampling error, or "bounce," that affects measures like GRPs, Reach, and Frequency for one survey period—the survey in which the commercials ran, for example. Those issues are discussed and quantified in *Section One* of this report.
- ▶ **Forecasting Error:** Evaluating the audience estimate for a commercial schedule involves comparing a survey's results to a user's expectation. Sometimes that expectation is based on negotiated predictions of future performance; similarly, it can reflect accepted patterns of difference across time (e.g., seasonality or sports programming differences). The potential sources of error in such prediction are many, and we have not attempted to quantify them here, though we have recently published a companion piece on radio seasonality called *Radio Year-Round*. However, we did study one particular form of prediction, discussed in the next item.
- ▶ **Sampling Error when Comparing Two Surveys:** In some cases, a user's expectation of a schedule's audience may be shaped purely by a prior survey's results. For example, the user may expect that the GRP estimates for the actual schedule will be identical to the results seen in a particular prior survey. In that case, there are two sources of sampling error—the survey used for planning and the survey in which the spots actually ran. In *Section Two* of this report, we provide the tools to estimate that combined effect.
- ▶ **Variance from Comparing Hour-Based GRPs to Daypart-Based Expectations:** This issue is a part of what's sometimes referred to as, "buying dayparts, but posting hours." It refers to the possibility that some users may expect an hour-based GRP to be equal to a daypart-based GRP because they expected "even rotation" through a daypart. When the number of spots is equal to an even multiple of the number of hours, there's no *statistical* issue with that practice; the GRP sum and its sampling error would be equal under either computation scenario.

However, when the number of spots is *not* equal to an even multiple of the number of hours, it's likely that an hour-based GRP will differ from a daypart-based GRP despite the delivery of "even rotation." That phenomenon is quantified in *Section Three* of this report.

- ▶ **Variance from Rounding in GRP Computation:** For GRPs in particular, there can be an additional source of numerical variance—namely, rounding—if certain computational methods are followed. Specifically, if the user simply adds together the relatively-rounded AQH rating points provided by Arbitron, there is an additional source of "bounce" which we've quantified in *Section Four* of this report.

Those issues describe the statistical and computational sources of variance we were able to isolate for this report. Users should consider as many of them as apply to a particular analytical situation. Of course, there are many other potential sources of "error" which cannot be so precisely quantified (see Appendix B).

Overall Summary of Results

Among the many findings in this report are the following important conclusions.

- ▶ **The reliability of a multiple-station GRP will be significantly better than the reliability of a single station's GRP.** Arbitron's estimate of radio's *total* delivery of audience to an advertiser can be a very reliable number; in some simulated-schedule tests of our model, we saw a 95% confidence interval for GRPs of less than $\pm 15\%$ for a three-station Monday-Friday 6AM-7PM buy in top-50 markets; it was as low as $\pm 4\text{-}5\%$ for a three-station broader-demo buy in a top-5 market. But a *single station's* estimated audience delivery can have a confidence interval two-to-three times as large as that of the campaign total.
- ▶ **The reliability of radio schedule audience delivery can vary significantly by market size.** In testing our model, we found that the #50-ranked market may have a confidence interval around schedule audience estimates that's twice as large as in New York.
- ▶ **The breadth of age/sex demographics affects the reliability of radio schedule audience estimates, though narrower demos are more reliable than their sample sizes might suggest.** In some of our evaluation markets, Men 18-34 represented only a fourth as much sample as Adults 18-49, but the confidence interval for Men 18-34 schedules was only 50-70% larger.

- ▶ **Compared to GRPs, Reach and Frequency can be equally reliable estimates of radio schedule audience delivery. In some cases, they can be even slightly more reliable.** The reliability relationship between GRPs and Reach & Frequency will vary depending on the schedule, but tests of our model on simulated schedules showed striking parity between those two approaches.
- ▶ **If a prior survey's reliability is relevant—for example, if the reliability of the survey used for planning is also considered to be an issue—the size of the confidence interval grows by about 40%.** Some researchers have argued that users must account for the reliability of both the planning/buying survey and of the survey in which spots ran. If the prior survey's results are the *only* factor considered in setting expectations for the survey in which spots run, then the relevant Standard Error is about 40% larger than for a single survey. Additional details and a more precise estimation formula are provided in Section Two, "Sampling Error When Comparing Two Surveys."
- ▶ **Despite the delivery of an "even rotation" by a station, GRPs computed from hourly ratings can differ meaningfully from daypart-based GRPs if the number of spots contributing to the GRP is low.** Because hourly ratings can differ from the daypart average, the placement of "odd spots"—those beyond an even multiple of the number of hours in a daypart—can cause additional variance from the daypart average rating. In a set of test data we analyzed, five spots run evenly in the four-hour daypart Morning Drive had an hour-based GRP that was at least 4.5% different from the daypart-based GRP about a third of the time.
 This effect diminished quickly, however, as the number of spots increased. For a majority of the situations analyzed, this effect dropped below 1% for nine or more spots in a four-hour daypart; it took 17 or more spots for the effect to drop below 1% for a broad daypart like Monday-Friday 6AM-7PM. Additional details are in Section Three, "Variance from Comparing Hour-Based GRPs with Daypart Expectations."
- ▶ **Computing GRPs from Arbitron's published AQH ratings (vs. unrounded persons projections) can add meaningful additional variance to that GRP.** If GRPs are computed from rating points, users should account for rounding as an additional source of error. That error varies depending on the number of different Arbitron ratings used in the computation, and on the sizes of the ratings. For example, if a GRP for a single station is computed from four different Arbitron ratings (e.g., using the four hourly ratings comprising Morning Drive), and those ratings average about 1.0 in magnitude, the additional error could be expressed in 95% confidence interval terms as plus-or-minus about 3 percent.
 If a user finds it necessary to compute GRPs by adding rating points, then the analysis presented in Section Four should be taken into account. That rounding error constitutes an additional, independent source of variance that could cause a GRP to differ from expectations in ways that can now be quantified.

The above findings are described in more detail in each of the Sections which follow.

Section One: Sampling Error Within One Survey

Study Scale: 30 Metros, 104,166 Diaries and 39,600 Schedules

To ensure a robust model, this primary portion of the study was very large, using the Arbitron syndicated diary samples from Winter 1995 for 30 selected top-50 Metros. The markets are listed in Section Five, and they ranged in size from New York to Memphis, representing 104,166 in-tab diaries.

We developed simulated spot schedules consisting of from one to five of the top stations in a Metro for each particular daypart and demo, based on AQH persons ranking. For reasons explained later, the resulting model for multiple stations will apply to other combinations of stations, beyond just the top stations. For each station and daypart, we constructed schedules representing three different spot volumes. These are the *total* number of spots per station in our simulations:

- ▶ One per hour (e.g., 4 spots per station for the Morning Drive-only schedules)
- ▶ Two per hour (e.g., 8 spots in Morning Drive, or 26 for M-F 6AM-7PM)
- ▶ Four per hour (e.g., 16 spots in Morning Drive, or 52 for M-F 6AM-7PM)

Because of the way GRPs are calculated (the sum of survey-average estimates), the number of weeks in the schedule is irrelevant; only the total number of spots in an hour-long daypart matters. The GRP total for 4 spots per week over 4 weeks is equal to (and would have the same reliability as) 16 spots in one week, all else being equal.

The same applies to Reach in this study, since we used single-week Reach estimates. In every case, we computed Reach (and Frequency) as if all spots ran in one week.

This simulation resulted in 39,600 different spot schedules. While real-world schedules can obviously fall outside these characteristics, this wide range of schedules provides ample variation for the purposes of this study—to develop a *model* of reliability variation that can apply to most all situations.

Other Variables

The study attempts to address whether, and to what extent, radio schedule audience estimate reliability differs for the following variables.

- Metro size (i.e., amount of in-tab)
- Dayparts (via selected dayparts, producing a number-of-quarter-hours proxy)
- Common age/sex demographics

Number of stations included in the schedule

Number of spots included in the schedule

Reliability Estimation: A Few Technical Definitions

In this study, we are interested in estimating the reliability of radio-schedule audience estimates. By “reliability,” we mean a measure of what’s often referred to as “bounce”—the stability of a measure from one sample to the next. This margin of error is known as sampling error, and is often quantified by the “Standard Error” measure.

The Standard Error can be used to calculate confidence intervals. For example, a 95% confidence interval for a GRP total is plus or minus 1.96 Standard Errors from the GRP value, and is interpreted as follows: If the survey were based on the population rather than a sample, the GRP value would fall within this confidence interval with a probability of 95%.

In real-world surveys, the actual size of the Standard Error depends on many factors, as previously documented in Arbitron research;¹ textbook “simple random sample” formulas are inadequate for Arbitron surveys because of factors such as sample design, multiple persons per household, and repeated measurement of persons over time. In this study, we used the Taylor series method to compute directly the Standard Error of the GRPs and Reach for each schedule; the Standard Error of Frequency is estimated from the GRP and Reach data. This method involves recomputing the relevant measures (e.g., GRP) at the in-tab household level, computing the actual standard deviations within that sample, and then estimating the average Standard Error through established statistical formulas. Additional details of the method will be provided upon request.

As a practical matter, it’s often more useful to discuss the “Relative Standard Error”—the Standard Error expressed as a *percentage* of the measure. For example, across all the 39,600 schedules in this study, the Relative Standard Error for GRPs was 7.6%. On average, for these schedules, the Standard Error was 7.6% of the GRP, and the 95% confidence interval would thus be plus or minus 14.9% (or, $\pm 1.96 \times 7.6\%$) of the GRP total.² We will be referring to the Relative Standard Error measure frequently in this report.

Finally, we will provide certain data in the form of a “Statistical Efficiency” value. It’s the Statistical Efficiency number which allows application of our findings to the simple in-tab numbers which appear in Arbitron reports. It will come as no surprise to most readers that the reliability of GRP, Reach, and Frequency numbers depends heavily on actual in-tab size for a particular market and demographic. The Statistical Efficiency number allows further adjustment

¹M. Occhiogrosso and M. Frankel (1982) *Arbitron Replication II: A Study of the Reliability of Radio Ratings*. Arbitron Ratings Company, Laurel, Maryland.

²The reader is cautioned against attributing too much meaning to this average of 14.9%, since we do not pretend that the “average schedule” in this study has any direct linkage to the real-world average of typical radio schedules. By design, our study includes a wide range of schedules which may not in fact occur very often in practice.

of an in-tab number into the reliability estimates appropriate to that sample. In other words, the Statistical Efficiency number accounts for all the *other* factors (besides rating size and simple sample size) which affects the reliability of a given estimate. These effects are sometimes referred to as “design effects.”

To put it another way, the Statistical Efficiency represents whether a particular Arbitron sample behaves statistically as if it were larger or smaller than a “simple random sample.” A Statistical Efficiency of 2.0, for example, means that the Arbitron sample actually behaves (in reliability terms) as if it were twice as large as the raw in-tab number. In the next several sections, we will attempt to explain *why* the patterns are the way they are. If you wish to skip the explanations for now and move quickly to the application of the findings, we refer you to the section titled, “Computing Single-Survey Confidence Intervals,” beginning on page 17.

Variables Affecting GRP Reliability

Obviously, the reliability of GRPs varies by sample size: The Relative Standard Error of a GRP total will vary in proportion to $1/\sqrt{IN-TAB}$. Within the range of in-tab sizes in this analysis, for example, the average Relative Standard Error of a GRP approximately doubles in size as one moves down the market list from New York (4.8%) to Memphis (10.1%).³ Clearly, market size is one of the most important variables affecting the reliability of GRPs.

In addition to that “given,” though, our study also finds that:

The reliability of GRPs varies substantially by demographic group. But narrower demos are more statistically efficient. In other words, a narrow demo may be less reliable than a broader, larger demo, but the loss of reliability is less than one might estimate from the in-tab numbers alone.

Actually, this isn’t a new finding. Much the same was found in *Replication II*. Nevertheless, it’s reassuring to see the pattern repeated in the current study. As explained there, the efficiency of a demo is related to both the amount of within-household respondent clustering affecting the demo, and the relative homogeneity of that demo’s listening preferences.

³Again, we caution you not to read too much into these two averages’ absolute values, for reasons mentioned earlier. What’s striking is the relative difference between them.

One way of looking at this is through the Statistical Efficiency measure discussed above. As demos get narrower, the persons-per-household gets smaller; for example, we found only 1.03 Men 35+ in-tab diaries per household in these samples, but there were 1.49 Adults 25-54 per household. That makes a Men 35+ sample more statistically efficient than Adults 25-54, which translates into a higher Statistical Efficiency for that group (8.0 vs. 6.3).

Also affecting this pattern is how much variation in listening occurs within a demographic group. Thus, Females 12-34 have a slightly higher Statistical Efficiency (9.4) than Men 35+ (8.8), even though there are more of the former per household (1.16) than for the latter (1.03). That sort of pattern is usually attributable to how much variation in listening occurs within a demo group; less variation in listening translates into higher efficiency.

The demos we considered in this study had GRP Statistical Efficiencies which ranged from 5.5 for Persons 12+ to 9.4 for Females 12-34. This variation in reliability is shown in the tables presented at the end of this section. Note the absolute size of the Statistical Efficiencies, too. This suggests that Arbitron's samples continue to have much more reliability than simple sample size suggests, even in this application.

The bottom line: The size of a demographic group has a significant relationship to the reliability of GRPs, potentially as important as market size. But the variation is less than one might expect from the differences in sample sizes alone.

The Relative Standard Error measure captures both the efficiency and the size of the sample. Here are just a few of the average GRP Relative Standard Errors we saw across all schedules and markets in our study:

Adults 18-49	6.7%
Adults 25-54	6.8%
Women 35+	7.2%
Men 35+	7.4%
Women 18-49	7.6%
Women 25-54	7.8%
Men 25-54	8.5%
Women 18-34	9.6%
Men 18-34	10.5%

The reliability of GRPs improves somewhat as the schedule is spread over more dayparts.

This finding, too, mirrors past research. *All else being equal*, a GRP total from a schedule which is spread across multiple dayparts will be more reliable than one which is concentrated in fewer dayparts. This is also true of AQH ratings, as explained in *Replication II*, and it reflects the reliability benefit of having more measures over time per in-tab person.

This effect is apparent in the Statistical Efficiency numbers. The Monday-Friday 6AM-10AM daypart, for example, has a GRP Statistical Efficiency of 5.1 across all of our schedules, while the broader Monday-Sunday 6AM-Midnight daypart has an efficiency of 12.1.

This variation by daypart is accounted for in the tables presented at the end of this section.

The Relative Standard Error of GRPs does not vary if the spot schedule is decreased or increased by the same factor in all the stations and dayparts. For example, doubling the number of spots *per se* has no impact on GRP Statistical Efficiency, and therefore, on relative error.

This may be one of the most surprising findings from our study, although in fact, one doesn't need a study like ours to come to this conclusion. It's just an algebraic fact that doubling or quadrupling the number of spots has no impact on the statistical efficiency or Relative Standard Error of a GRP (assuming all else is constant).

Note, however, that it's the *Relative* Standard Error that doesn't vary. For example, if the Relative Standard Error of a GRP of 100 is 10%, the 10% estimate would also hold for a GRP of 50 or 200, if all one changed was the number of spots (i.e., same stations, same dayparts, etc.). But the *absolute* Standard Error does vary, of course—anywhere from 5 to 20, in this example.

The simplest explanation for this phenomenon rests with the constant underpinnings of a GRP. A GRP is nothing more than some multiple of Arbitron's 12-week survey average AQH audience estimate, expressed as a rating. A GRP for one spot would have the same reliability as the AQH estimate that was applied to it. If two spots ran in the same hour, then we have simply multiplied that estimate by a constant; we've made no change to the underlying survey result, the AQH audience. Thus, if the AQH estimate is multiplied by two, its absolute Standard Error is multiplied by two, and the Relative Standard Error stays the same.

It might help to use a simple analogy. Suppose you buy a single share of stock, which has a certain volatility over time. Let's suppose we could quantify that volatility into something like a confidence interval, by saying that the stock is likely to vary in value by plus-or-minus ten percent over a three-month period. Thus, the amount you invested in that share of stock could rise or fall in value by $\pm 10\%$ per quarter.

Now let's suppose you've come into an inheritance, and you decide to buy 100 shares of that same stock. The volatility stays the same in percentage terms, because it's dependent on the swings in per-share price; your investment would still rise or fall in value by $\pm 10\%$. Of course, the absolute dollars represented by those swings would be substantially more (100 times as great, to be precise).

That's the same basic arithmetic behind the conclusion about GRPs. Note that this finding is true only of GRPs. As you will see shortly, Reach and Frequency are another story.

The reliability of a GRP total improves substantially as more stations are added to the mix. A total GRP for a multiple-station campaign will have a much smaller Relative Standard Error than will the GRPs for each station in the campaign.

Roughly speaking, this finding is the result of risk diversification. The sampling bounce in the ratings for one station tends to be either unrelated to, or even negatively correlated with, those of other stations.

The effect on the data is meaningful in practice, we think. For example, across all the single-station schedules in our analysis, the GRP Relative Standard Error was 11.1%, but for schedules in which the spots were spread across five stations, the Relative Standard Error was 5.7%. We also saw wide variations in the underlying Statistical Efficiency values.

To express it a different way: If one holds the GRP level constant, and simply spreads the same buy evenly across more stations, the resultant GRP becomes more reliable. A two-station buy would have a Standard Error about 30% smaller than a one-station buy; a three-station buy's Standard Error would be about 42% smaller; a four-station's would be 50% smaller; and a five-station's, about 55% smaller, all else being equal. The buy may or may not be more *effective* that way, since spreading the same number of spots across stations would tend to increase Reach at the expense of Frequency. But in *reliability* terms, there is clear improvement.

The importance of this factor varies according to the user's purpose. If the purpose is to evaluate the delivery of a single station, then this variable is irrelevant. But if the purpose is to estimate the delivery of a total campaign, then the number of stations contributing to a GRP is an important variable, which you will see in the formulas provided at the end of this section.

Note that the above examples of efficiency gains are based on equal numbers of spots per station. The efficiency gain would not be as dramatic if a large number of spots were placed on one station, and only a trivial number were placed on other stations. The formulas provided will account for differing numbers of spots per station.

Variables Affecting Net Reach Reliability

Another audience estimate associated with radio spot schedules is the Reach (or Net Reach) of the schedule. This is defined as the persons among the target group who were exposed to one or more spots of the schedule (which we express as a rating in our analysis). Unlike the GRP, Reach is an unduplicated audience.

Though conceptually simple, the laboriousness of precisely calculating Reach has led to a variety of conventions and estimating algorithms in actual practice. Here are the conventions we used for the current analysis:

- ▶ **One-Week Reach:** Because of the proliferation of multi-week estimation models, we decided not to burden the current study with the additional problem of having to defend the choice of a particular multi-week estimation model. The Reach reliability patterns we discover should apply to other mathematically-similar models.
- ▶ **Book Averages:** Even if one is looking at one-week Reach, it is conventional to use the one-week Reach that is averaged across all 12 weeks of the survey. This convention is helpful for reliability, as the user then benefits from the use of all 12 weeks of sample in the survey.
- ▶ **Placement Issues:** As a matter of computational expediency, Reach is often calculated without regard for the exact placement of spots within a daypart; rather, the average Reach is calculated from all possible evenly-spread spot placements within a daypart. Here, we use the hypergeometric estimation model used within Arbitron's Maximi\$er product, wherein each respondent is given a probability of exposure for the number of spots that ran within a standard daypart.

With those caveats, here's what we learned about the reliability of Reach.

Again, market size is one of the greatest variables affecting Reach reliability. Across the wide range of schedules and demos used for this analysis (which, again, may not represent the "typical" buy), we saw an average Relative Standard Error for Reach of 4.9%. That average ranged from 3.4% in New York to 6.6% in Memphis. That's to be expected. In addition to market size findings:

As was true for GRPs, the choice of demographic group has a significant effect on reliability. Smaller demos have less reliable Reach estimates, but the loss is less than simple formulas predict.

The obvious effects of sample size are ameliorated by variations in household composition and listening patterns, as we saw for GRPs. Here are some of the variations in Relative Standard Error that we saw across demos:

Adults 18-49	4.4%
Adults 25-54	4.5%
Women 35+	4.8%
Men 35+	4.8%
Women 18-49	4.9%
Women 25-54	5.0%
Men 25-54	5.4%
Women 18-34	6.0%
Men 18-34	6.4%

Note again how a demo like Men 18-34, which had only 23% of the sample size of the Adults 18-49 demo, sees only a 45% increase in its Relative Standard Error. Narrow demos are less reliable, but the margin of difference is smaller than the decrease in sample size would seem to imply.

Demographic variation is an important part of the tables presented in later sections.

The relative reliability of Reach improves somewhat as the schedule is spread over more dayparts—despite a decline in sample efficiency.

Since Reach is in part a cume-type measure, it doesn't benefit from multiple measures per person the way AQH and GRP can. Nevertheless, Reach does have some unique variation in Efficiency that correlates with the breadth of the daypart.

Reach (as we calculated it) is actually the average *probability* of exposure to a schedule. Thus, its statistical behavior can be somewhat different from the simple binomial cume measure. For narrower dayparts, we found that the use of Reach probabilities makes the sample behave more efficiently than for a cume measure. For broader dayparts or heavier schedules, the respondent-

level probabilities of exposure tend toward the extremes of 100% or zero, and in those special cases, Reach begins to behave more like a cume rating (in reliability terms). That translates into a slight loss of efficiency as the daypart gets larger.

Overall, daypart is a relatively small factor for Reach reliability. Nevertheless, daypart variation is a part of the tables presented later.

Unlike GRPs, the reliability of Reach does improve somewhat when the number of spots is increased. But the effect is a relatively small one.

While this is still not a major factor, even for Reach, there is some evidence in our study of an improvement in Relative Standard Error for heavier schedules, all else being equal. But this is driven by the tendency toward higher Reach values, not by any improvement in underlying Statistical Efficiency.

For example, our lightest schedules had an average Relative Standard Error for Reach of 5.2%; the heaviest schedules showed a reduction of that Standard Error to 4.6%. Overall, though, this factor pales in comparison to the others, and is not one of the variables in our look-up table; its effect is accounted for in the formula provided later, which takes rating size into account.

The reliability of Reach can improve somewhat as more stations are added to the mix. But that's mostly a result of increasing the size of the Reach estimate, not of increased statistical efficiency.

Within the range of demos, dayparts, and schedules we considered, the average Standard Error for a single-station Reach was 7.4%, while the five-station schedules averaged a more reliable 3.5%.

However, there's actually very little difference in the underlying Statistical Efficiency values, and we don't include that factor in the model for Reach. It appears that this relationship of number of stations and Reach reliability is driven more by the size of the resulting Reach estimate than by the efficiency of adding more stations, and is adequately accounted for in the formulas we provide.

Variables Affecting Frequency Reliability

Another audience estimate associated with radio spot schedules is Frequency. This is defined as the Gross Rating Points divided by the Reach rating, and is interpreted as the average number of spots in the schedule to which the unduplicated audience was exposed.

In our case, Frequency was computed after the GRPs and Net Reach were computed, and the reliability measures were derived from those of Frequency's components. This means that Frequency is subject to the constraints of the conventions that were adopted in the calculation of the other two quantities.

The range of schedules we considered here gave us a wide variety of Frequency levels, ranging from 1.3 to a high of 32.7. So this study should be robust enough to apply to all real-world situations.

At the risk of again stating the obvious, market size (in-tab size) has a powerful relationship to the reliability of Frequency estimates. Across all the demos, dayparts, and schedules in our study, the average Frequency estimate in New York had a Relative Standard Error of 3.2%, while the smallest market studied, Memphis, showed a bit over twice the Standard Error at 6.9%.

Aside from the effect of total market size, we also saw the following:

Once again, the choice of demographic group has a significant effect on reliability for Frequency. But the effect of varying sample sizes is smoothed somewhat by the efficiency benefits of narrower targets.

The obvious effects of sample size are again ameliorated by variations in household composition and listening patterns, as we saw for GRPs and Reach. Here are some of the variations in Relative Standard Error we saw across demos:

Adults 18-49	4.4%
Adults 25-54	4.5%
Women 35+	4.8%
Men 35+	5.0%
Women 18-49	5.2%
Women 25-54	5.2%
Men 25-54	5.8%

Women 18-34	6.6%
Men 18-34	7.3%

Once again, though Men 18-34 were only a fourth of the sample size of the Adults 18-49 demo, its Relative Standard Error grows by only 66%. Narrow demos are less reliable, but the margin of difference is smaller than some would expect.

Demographic variation for Frequency will be accounted for in the formulas we provide.

The reliability of Frequency varies only slightly as the schedule is spread over more dayparts.

Perhaps because of the relatively small variation we saw by daypart for Reach reliability, the variation for Frequency changed only slightly and somewhat unpredictably. For example, the Relative Standard Error for Monday-Friday 6AM-10AM averaged 4.1% across all of our schedules, slightly less than the total week Monday-Sunday 6AM-Midnight average of 5.9%.

Overall, it appears that daypart selection is relatively unimportant to estimating the reliability of Frequency estimates, and it will be adequately accounted for in the formulas.

The reliability of Frequency does *not* improve when the number of spots is increased. In fact, for the schedules studied, there was a small tendency toward *reduced* reliability.

Within the range of schedules studied here, we actually detected a slight worsening of the reliability of Frequency as the number of spots increased toward the maximum. For example, the average Standard Error for our lightest schedules was 4.5%; that figure rose to 5.7% for the heaviest schedules.

Overall, though, it appears that number of spots has a relatively small effect on reliability for Frequency, perhaps because of the combined effect of its irrelevance to GRPs and its small importance to Reach.

As with GRPs and Reach, the reliability of Frequency improves as more stations are added to the mix. A total Frequency for a multiple-station campaign will have a smaller Relative Standard Error than will the Frequency for each station in the campaign.

As we would expect because of the behavior of its component parts, the reliability of Frequency estimates increases as more stations are added to the schedule.

Within the range of demos, dayparts, and schedules we considered, the average Standard Error for a single-station Frequency was 7.2%, while the five-station schedules averaged a more reliable 4.0%. This is similar to the difference seen with GRPs and Reach, so it's not surprising that a measure derived from them would behave in a like manner.

Computing Single-Survey Confidence Intervals

This section provides tables of Statistical Efficiency values and formulas that can be used to estimate the reliability of any real-world radio schedule's delivery, accounting for the sampling error of a single survey. Combined with the user's knowledge of...

- ▶ the in-tab sample size for the demographic and market involved,
- ▶ the broadest daypart that represents the schedule,
- ▶ the number of spots in the campaign,
- ▶ and the number of stations in the campaign,

...the Statistical Efficiency and formulas below will allow the computation of an estimated Standard Error for that GRP, Reach, or Frequency measure for a single survey.

Once the Standard Error is computed, the user should then compute a confidence interval appropriate to the application. For example, multiplying the Standard Error by 1.96 provides a plus-or-minus value which yields the 95% confidence interval.

Here are the specific methods for each of our three main measures of radio schedule audience estimate delivery.

Estimating Single-Survey GRP Reliability (Confidence Intervals)

Here's the formula that applies to GRPs for a single station in a single survey.⁴

The single-survey Standard Error of a single station's GRP total would equal:

$$\sqrt{(\text{GRP} \times ((100 \times \text{\#spots}) - \text{GRP}) \div (\text{In-tab} \times \text{Statistical Efficiency from Table 1.0}))}$$

For this formula, "GRP" represents the total GRPs delivered by a particular station. The term "#spots" is the total number of spots that ran on that station. And the term "in-tab" is the

⁴Details on the derivation of the formulas here and in later sections are available on request.

number of diaries that were in tabulation (returned and usable) for that market, survey and demographic.

For the evaluation of a multi-station campaign, the formula becomes slightly more complex, and requires the look-up of one additional number from Table 1.1, "Multiple-Station Adjustment Factor (MS) for GRPs."

For a multi-station evaluation, the single-survey Standard Error of the total campaign's GRP would equal:

$$\sqrt{(\text{Sum of } \text{GRP}_{\text{station}} \times ((100 \times \# \text{spots}_{\text{station}}) - \text{GRP}_{\text{station}}) \div (\text{In-tab} \times \text{Stat. Efficiency} \times \text{MS Factor}))}$$

The key differences with the multiple-station version of the formula are:

- ▶ We've added a further adjustment to the Statistical Efficiency (the MS Factor) that accounts for some variations in the statistical independence of station ratings for different demos and numbers of stations;⁵
- ▶ And we're computing the following factor *for each station*, and adding the results together, *before* taking the square root:

$$(\text{GRP} \times ((100 \times \# \text{spots}) - \text{GRP}) \div (\text{In-tab} \times \text{Statistical Efficiency} \times \text{MS Factor}))$$

Participants in our planning discussions may remember that we had originally intended to provide separate look-up table values that corresponded to the number of stations. But those values would have required an assumption of equal numbers of spots across stations, an unrealistic assumption. The formula above is a more general approach, which will apply to unequal numbers of spots across stations.

Now, let's take a specific *single-station* example and walk through the process step-by-step:

In-Tab Size:	3,049 diaries in-tab
Demographic:	Women 25-54
Daypart:	Monday-Friday 6AM-7PM
Number of Stations:	1
Number of Spots:	156 (13 per week × 12 weeks)
GRPs Delivered:	418.8 (34.9 per week)

Here's how the reliability of that single-station campaign is estimated for GRPs:

⁵More precisely, we've added an adjustment factor to account for the slight shifts in the covariance of GRPs between stations. That covariance is near zero most of the time, and even skews slightly negative, which is what allows for the multiple-station efficiency gains discussed earlier. We've allowed for slight variations in that relationship by integrating the MS factor in our formula, based on the observed covariances.

- 1) **Determine the daypart which is the closest match to the overall schedule, and refer to the column on Table 1.0 that fits that daypart best.**

The columns are labeled in terms of total quarter-hours per week within that daypart, since it's the number of "repeated measures" that really matters for statistical efficiency. The example daypart (Monday-Friday 6AM-7PM) represents 260 quarter-hours, so the third column from the right is the appropriate column.

- 2) **Determine which of the 22 listed demographic groups is the closest match to the one that is used in this spot schedule.**

In this case, the demographic under evaluation, Women 25-54, appears on the table.

For a demographic that does not appear on the table, we recommend using the same-gender demo (Men, Women, Adults) that's closest in age range and which completely encompasses the desired demo. For example, if the user wished to estimate the Statistical Efficiency for Women 25-64, which does not appear on the table, we would recommend using Women 18+ values from the table. This approach is conservative, in that it will slightly overstate the Standard Error (by using a slightly understated Efficiency).

- 3) **From the row corresponding to that demographic, select the Table value (cell) which corresponds to the number-of-quarter-hours column selected in Step 1. That's the best estimate of Statistical Efficiency for this schedule.**

In our example, you can see that the Statistical Efficiency for Women 25-54 and 260 Quarter Hours is equal to 2.93.

- 4) **Compute the (absolute) Standard Error using the single-station formula:**

$$\sqrt{(GRP \times ((100 \times \#spots) - GRP) \div (In-tab \times \text{Statistical Efficiency}))}$$

In our example, we'd insert GRP=418.8, # of spots=156, In-Tab=3,049, and Statistical Efficiency (from steps 1-3)=2.93. The formula would then look like:

$$\sqrt{(418.8 \times ((100 \times 156) - 156) \div (3049 \times 2.93))}$$

which equals 26.9. That's the absolute Standard Error; the Relative Standard Error is expressed as a percentage (26.9x100÷418.8), which equals 6.4%.

- 5) **Calculate the appropriate confidence interval using the usual multiples of Standard Errors:**

±1.00 Standard Error	=	68% confidence interval
±1.65 Standard Errors	=	90% confidence interval

±1.96 Standard Errors = 95% confidence interval
±2.58 Standard Errors = 99% confidence interval

As previously noted, the confidence interval is interpreted to mean that if the survey were based on the population rather than a sample, the GRP value would fall within the confidence interval with a probability of XX%. Which confidence interval one chooses depends on many factors, including one's tolerance for risk and the consequences of an incorrect conclusion.

In our example, let's suppose the user wished to know the 95% confidence interval around the GRP estimate. Since we estimated that the Standard Error is 26.9 GRPs, the 95% confidence interval equals plus-or-minus 1.96×26.9 , or ± 52.7 .

Thus, we could say that the campaign on this station delivered 418.8 GRPs, with a 95% confidence interval of plus or minus 52.7 points (or $\pm 12.6\%$). A user could be 95% confident that a survey of the total population (rather than a sample) would have shown a delivery somewhere between 366.1 and 471.5 GRPs.

The sequence of steps for evaluating a multiple-station GRP would be similar. There would be the additional step of looking up the MS Factor to adjust each Statistical Efficiency, and the order of computation is slightly different, as previously described.

Estimating Single-Survey Reach Reliability (Confidence Intervals)

The process of estimating reliability for specific Reach estimates is similar, though not identical, to the process for single-station GRPs. Because we have established that Number of Stations is not relevant to Statistical Efficiency for Reach, we provide only one formula, and all Reach-related values are included on one table, Table 2.0.

Thus, the user only needs to know the Net Reach delivered by the campaign (expressed as a rating), the in-tab size, the demographic group, and the daypart.

The single-survey Standard Error of a Reach total equals:

$$\sqrt{((\text{Reach} \times (100 - \text{Reach})) + (\text{In-Tab} \times \text{Statistical Efficiency}))}$$

We'll illustrate each step with the following data:

In-Tab Size:	3,049 diaries in-tab
Demographic:	Women 25-54

Daypart:	Monday-Friday 6AM-7PM
Number of Stations:	N/A
Number of Spots:	N/A
Reach Delivered:	10.9%

The steps for estimating a confidence interval for a Reach estimate are as follows:

- 1) Determine the daypart which is the closest match to the overall schedule, and refer to the column in Table 2.0 that fits that daypart best.**

The columns are labeled in terms of total quarter-hours per week within that daypart, since it's the breadth of the daypart that really matters for statistical efficiency. The example daypart (Monday-Friday 6AM-7PM) represents 260 quarter-hours, so the sixth column from Table 2.0 is the appropriate column.

- 2) Determine which of the 22 listed demographic groups is the closest match to the one that is used in this spot schedule. From the row corresponding to that demographic, select the Table value (cell) which corresponds to the number of quarter-hours in the schedule. That's the best estimate of Statistical Efficiency for Reach for this schedule.**

In this case, the demographic under evaluation, Women 25-54, appears on the table.

For a demographic that does not appear on the table, we recommend using the same-gender demo (Men, Women, Adults) that's closest in age range and which completely encompasses the desired demo. For example, if the user wished to estimate the Statistical Efficiency for Women 25-64, which does not appear on the table, we would recommend using Women 18+ values from the table. This approach is conservative, in that it will slightly overstate the Standard Error (by using a slightly understated Efficiency).

In our example, the value we seek is in the sixth column. On Table 2.0, you can see that the Statistical Efficiency for Women 25-54 is equal to 1.16.

- 3) Compute the (absolute) Standard Error using the formula:**

$$\sqrt{((\text{Reach} \times (100 - \text{Reach})) \div (\text{In-Tab} \times \text{Statistical Efficiency}))}$$

For our example, this formula would translate into:

$$\sqrt{((10.9 \times (100 - 10.9)) \div (3,049 \times 1.16))} = 0.524$$

That's the absolute Standard Error; the Relative Standard Error is expressed as a percentage (0.524 as a percent of 10.9), or 4.8%.

- 4) Calculate the appropriate confidence interval using the usual multiples of Standard Errors:**

±1.00 Standard Error	=	68% confidence interval
±1.65 Standard Errors	=	90% confidence interval
±1.96 Standard Errors	=	95% confidence interval
±2.58 Standard Errors	=	99% confidence interval

As previously noted, the confidence interval is interpreted to mean, if the survey were based on the population rather than a sample, the Reach value would fall within the confidence interval with a probability of XX%. Which confidence interval one chooses depends on many factors, including one's tolerance for risk and the consequences of an incorrect conclusion.

In our example, let's suppose the user wished to know the 95% confidence interval around the Reach estimate. Since we estimated that the absolute Standard Error was 0.524 points, the 95% confidence interval would equal plus-or-minus 1.96×0.524 , or ± 1.03 points.

Thus, we could say that the campaign on this station delivered a Reach of 10.9% with a 95% confidence interval of plus or minus 1.0 points (or $\pm 9.4\%$). A user could be 95% confident that a survey of the total population (rather than a sample) would have shown a Reach delivery somewhere between 9.9% and 11.9%.

Estimating Single-Survey Frequency Reliability (Confidence Intervals)

Because the Standard Errors and confidence intervals for Frequency are derived from the data for GRPs and Reach, no further look-up tables are provided. Rather, the computation proceeds directly to another pair of formulas. In the examples provided, we'll continue with the Women 25-54 example developed above, for which the Frequency was calculated to be 3.2.

- 1) First, compute the Relative Standard Error (RSE) of Frequency by applying the formula below to the Relative Standard Errors of the corresponding GRP and Reach estimates:**

$$RSE_{(freq)} = \sqrt{(RSE_{(GRP)}^2 + RSE_{(Reach)}^2 - (2 \times RSE_{(GRP)} \times RSE_{(Reach)} \times 0.75))}$$

For those who are familiar with statistics, that last number (0.75) is actually the Correlation Coefficient of GRP and Reach. We discovered that the Coefficient was very stable in this study, with an average value of 0.75. That's fortunate, since a simple estimation method for Frequency might not be possible without that shortcut.

To continue with our examples from earlier: The Relative Standard Error of GRP in our example was equal to 6.4% (the Standard Error of 26.9 as a percentage of the GRP

value of 418.8). The Relative Standard Error of Reach for our example was shown earlier to be 4.2% (0.52 as a percent of 10.9). Therefore, the Frequency calculation would appear as follows:

$$RSE_{(freq)} = \sqrt{(6.4^2 + 4.2^2 - (2 \times 6.4 \times 4.2 \times 0.75))} = 4.2\%$$

2) Calculate the appropriate confidence interval using the usual multiples of Standard Errors:

±1.00 Standard Error	=	68% confidence interval
±1.65 Standard Errors	=	90% confidence interval
±1.96 Standard Errors	=	95% confidence interval
±2.58 Standard Errors	=	99% confidence interval

In our example, let's suppose the user wished to know the 95% confidence interval around the Frequency estimate. Since we estimate that the Relative Standard Error is 4.2% of the Frequency (which equals 3.2), the 95% confidence interval equals plus-or-minus $1.96 \times (.042 \times 3.2)$, or ± 0.26 .

Thus, we could say that the campaign on this station delivered a Frequency of 3.2, with a 95% confidence interval of plus or minus 0.26 points (or $\pm 8.1\%$). A user can be 95% confident that a survey of the total population would show a delivery somewhere between 2.94 and 3.46.

Table 1.0
Statistical Efficiency for Single-Station Gross Rating Points

Number of Quarter-Hours in Daypart								
DEMO GROUP	16-20	80	100	144	160	260	360	504
Persons 12+	1.44	1.65	1.72	1.86	1.93	2.37	3.03	3.68
Adults 18+	1.51	1.74	1.81	1.97	2.04	2.42	3.16	4.01
Men 18+	1.94	2.23	2.32	2.52	2.61	3.04	4.04	5.84
Women 18+	1.96	2.25	2.34	2.54	2.63	3.07	4.07	5.55
Persons 12-34	2.12	2.44	2.54	2.76	2.85	3.19	3.80	4.71
Males 12-34	2.27	2.61	2.71	2.95	3.05	3.31	4.29	5.85
Females 12-34	2.72	3.13	3.26	3.54	3.66	4.14	4.94	6.05
Adults 18-34	1.98	2.28	2.37	2.58	2.67	2.90	3.61	4.74
Men 18-34	2.01	2.31	2.40	2.61	2.70	2.96	3.95	5.83
Women 18-34	2.44	2.80	2.91	3.16	3.28	3.35	4.31	6.17
Adults 18-49	1.67	1.92	2.00	2.17	2.25	2.42	3.16	4.45
Men 18-49	1.87	2.15	2.24	2.43	2.52	2.79	3.75	5.65
Women 18-49	2.21	2.54	2.64	2.87	2.97	3.01	4.02	6.31
Adults 25-54	1.72	1.98	2.06	2.24	2.32	2.51	3.29	4.77
Men 25-54	2.04	2.34	2.43	2.64	2.74	3.03	4.06	6.40
Women 25-54	2.22	2.55	2.65	2.88	2.98	2.93	3.98	6.48
Adults 35-64	1.78	2.05	2.13	2.32	2.40	2.88	3.86	5.47
Men 35-64	2.17	2.49	2.59	2.81	2.91	3.24	4.35	6.77
Women 35-64	2.09	2.40	2.50	2.71	2.81	2.89	3.92	6.22
Adults 35+	1.64	1.89	1.97	2.14	2.21	2.91	3.81	4.85
Men 35+	2.24	2.58	2.68	2.92	3.02	3.79	5.01	7.34
Women 35+	2.03	2.33	2.42	2.63	2.73	3.34	4.41	5.98

Table 1.1
Multiple-Station Adjustment Factor (MS) For GRPs

DEMO GROUP	Number of Stations				
	1	2	3	4	5
Persons 12+	1.00	1.02	1.04	1.05	1.05
Adults 18+	1.00	1.02	1.05	1.06	1.07
Men 18+	1.00	1.04	1.04	1.05	1.07
Women 18+	1.00	1.00	1.02	1.04	1.07
Persons 12-34	1.00	1.03	1.04	1.04	1.04
Males 12-34	1.00	0.98	1.02	1.01	1.01
Females 12-34	1.00	0.99	1.00	1.00	1.01
Adults 18-34	1.00	1.05	1.08	1.09	1.08
Men 18-34	1.00	1.02	1.06	1.07	1.09
Women 18-34	1.00	1.07	1.06	1.06	1.09
Adults 18-49	1.00	1.02	1.04	1.06	1.06
Men 18-49	1.00	1.02	1.06	1.08	1.10
Women 18-49	1.00	0.99	1.01	1.03	1.06
Adults 25-54	1.00	1.00	1.03	1.05	1.05
Men 25-54	1.00	1.00	1.04	1.06	1.07
Women 25-54	1.00	1.03	1.05	1.07	1.10
Adults 35-64	1.00	0.97	0.96	0.99	1.00
Men 35-64	1.00	1.03	1.06	1.08	1.10
Women 35-64	1.00	1.06	1.08	1.12	1.16
Adults 35+	1.00	0.98	0.97	0.99	0.99
Men 35+	1.00	0.97	0.96	0.97	0.98
Women 35+	1.00	1.00	1.03	1.06	1.07

Table 2.0
Statistical Efficiency for Net Reach

Number of Quarter-Hours in Daypart

DEMO GROUP	16-20	80	100	144	160	260	360	504
Persons 12+	0.85	0.81	0.80	0.77	0.76	0.69	0.67	0.54
Adults 18+	0.91	0.87	0.86	0.83	0.82	0.76	0.74	0.60
Men 18+	1.21	1.16	1.15	1.11	1.10	1.03	1.01	0.83
Women 18+	1.20	1.16	1.15	1.12	1.11	1.04	1.02	0.84
Persons 12-34	1.07	1.00	0.98	0.93	0.91	0.79	0.74	0.55
Males 12-34	1.29	1.22	1.20	1.15	1.14	1.03	0.98	0.76
Females 12-34	1.32	1.24	1.22	1.16	1.14	1.02	0.97	0.73
Adults 18-34	1.12	1.05	1.03	0.98	0.97	0.86	0.83	0.63
Men 18-34	1.28	1.23	1.21	1.18	1.17	1.09	1.05	0.84
Women 18-34	1.36	1.30	1.28	1.24	1.22	1.12	1.09	0.86
Adults 18-49	1.00	0.95	0.93	0.90	0.89	0.81	0.78	0.62
Men 18-49	1.24	1.19	1.17	1.14	1.13	1.05	1.03	0.84
Women 18-49	1.29	1.23	1.21	1.17	1.16	1.07	1.04	0.83
Adults 25-54	1.03	0.99	0.98	0.95	0.94	0.87	0.85	0.68
Men 25-54	1.30	1.26	1.25	1.22	1.21	1.14	1.13	0.94
Women 25-54	1.35	1.30	1.28	1.25	1.24	1.16	1.14	0.94
Adults 35-64	1.05	1.00	0.99	0.95	0.94	0.87	0.86	0.69
Men 35-64	1.34	1.30	1.29	1.26	1.25	1.18	1.17	0.98
Women 35-64	1.34	1.30	1.29	1.26	1.25	1.18	1.17	0.98
Adults 35+	0.97	0.93	0.92	0.89	0.88	0.81	0.80	0.65
Men 35+	1.33	1.28	1.27	1.23	1.22	1.15	1.14	0.96
Women 35+	1.27	1.23	1.22	1.19	1.18	1.11	1.09	0.92

Section Two: Sampling Error When Comparing Two Surveys

During the pre-specification dialogue with customers, there was considerable debate among end users over whether the reliability issues are limited to the survey in which the spots ran. A number of researchers believe that users must also account for the reliability of the earlier survey used for planning.

The debate hinges on whether one believes that radio planning and buying involves de facto “forecasting” or not. If the buyer and/or seller are only using past data as a loose guide, and are in fact making a *prediction* about the delivery of the future schedule, perhaps based on changes in format, programming, or season, then the reliability of the older data is relatively unimportant. But if one believes that the older survey *is* the prediction—that the older survey’s results are what the buyer is expecting in the future—then some researchers believe that the older survey’s reliability is also relevant.

Arbitron is not in a position to resolve that debate; the question of which approach is “proper” must be left to the marketplace. However, we can at least quantify the additional error introduced when comparing one survey’s results to another.

In general, this comparison approach would require a difference between the two surveys *approximately* 40% larger than the single-survey’s confidence interval before concluding that the difference exceeded the chosen confidence interval. For example, if the 95% confidence interval for a single survey is $\pm 10\%$, the two-survey-comparison approach yields about a $\pm 14\%$ interval; the new survey would have to differ from the older survey by 14% or more before the user would conclude that the two differed with 95% confidence.

Statistically, this situation becomes a significance test of a measure’s difference between two independent samples. Rather than simply computing the Standard Error of a single GRP, for example, and using that number to construct a confidence interval, this alternative approach requires the computation of a value called the Standard Error of the Difference, and the computation of a confidence interval around the *difference* in GRPs.

Here’s an example. Under the single-survey scenario, the user computes the Standard Error for a particular GRP. Let’s take one of our earlier examples—a GRP of 27.2 with a Standard Error of 5.1% and a 95% confidence interval of $\pm 10.0\%$. In this scenario, the user simply decides whether that survey’s GRP is close enough to the number he or she was expecting—let’s say, a GRP of 30. Since the expected value of 30 is outside the 95% confidence interval for that survey’s GRP, the user would conclude that he or she was 95% confident that the actual GRP delivery was less than the expected value of 30; only one time in twenty would there be a difference of that size (30 vs. 27.2) from sampling error alone.

But what if the expected value of 30 is based purely on a separate, independent sample's results? What if the buyer and seller agree that no "real" change should occur between surveys, and that the listening patterns of the planning survey are the best estimates of the future? That planning survey, too, has a Standard Error of approximately equal size, and a confidence interval around it of approximately equal size. What now?

The Two-Survey Formula

The specific solution lies in a significance test of the difference, using the following steps:

- 1) First, compute the (absolute) Standard Error for each survey's measure (e.g., the standard error of survey #1's GRP and the standard error of survey #2's GRP), using the formulas elsewhere in this report.**

In the example above, the absolute Standard Error for survey #2 would be 1.53. The absolute Standard Error for the earlier survey #1 would be similar in size, assuming identical sample sizes; its value would be 1.61 in this case.

- 2) Compute the "Standard Error of the Difference," using the formula:**

$$\text{SED} = \sqrt{((\text{Standard Error \#1})^2 + (\text{Standard Error \#2})^2)}$$

For our example, this formula would become:

$$\text{SED} = \sqrt{((1.61)^2 + (1.53)^2)} = 2.22$$

- 3) Compute the "Z Score": Divide the difference between measures (e.g., the difference between GRPs) by the Standard Error of the Difference, as in:**

$$Z = ((\text{Measure \#1}) - (\text{Measure \#2})) \div \text{Standard Error of Difference}$$

For our example, this formula would become:

$$Z = (30 - 27.2) \div 2.22 = 1.26$$

The "Z Score" simply expresses the difference between two numbers as a multiple of the Standard Error of the Difference. In our example, then, we can say that the difference between the two GRPs is equal to 1.26 Standard Errors.

- 4) Compare this "Z score" to the confidence interval values shown below:**

If Z is larger than ± 1.00 then the difference exceeds the 68% confidence interval

If Z is larger than ± 1.65 then the difference exceeds the 90% confidence interval

If Z is larger than ± 1.96 then the difference exceeds the 95% confidence interval

If Z is larger than ± 2.58 then the difference exceeds the 99% confidence interval

In our example, in which Z equals 1.26, we can no longer say that the difference is large enough to achieve 95% confidence; it would have taken a Z value of 1.96 or larger.

Again, Arbitron is unable to recommend one approach over another. If one believes that buyers simply expect the same audience in a future survey as was achieved in one prior survey, then the difference-between-surveys approach outlined above seems appropriate. But once the buyers and sellers engage in any additional adjustment—for seasons, for changes in programming or competition, etc.—then additional variables have been introduced which no survey-to-survey formula can capture in any practical way.

Note that, in parallel with the current study, Arbitron is providing an updated analysis of radio audience seasonality, a new and enhanced edition of *Radio Year-Round*. That study, too, is the largest and most definitive yet, and should assist in the discussion of “prediction.”

Section Three: Variance from Comparing Hour-Based GRPs with Daypart Expectations

The Issue

One of the most contentious recent debates has been over what's known as "buying dayparts, but posting hours." Some agencies wish to conduct their schedule analysis on the basis of Arbitron's hour-by-hour data. Broadcasters believe that to be unfair if "the buy" was based on daypart planning and negotiating.

While the debate involves many, mostly nonstatistical, issues, Arbitron identified one component which we attempt to quantify here. Specifically, we studied schedules which contain a number of spots which are not equal to a multiple of the number of hours in the daypart.

An example: If the buyer buys four spots in a four-hour daypart and contracts for and receives equal rotation, there is no special reliability issue; the GRPs for those four spots should be the same (and have the same reliability) regardless of whether they're built from the use of daypart estimates for each spot or from the use of hourly estimates for each spot.

However, if the buyer purchased *five* spots in a four-hour daypart, the delivery estimate *could* vary depending on whether hourly or daypart estimates are attributed to each spot. Even under truly even rotation, those odd spots could easily fall into hours with audiences that are above- or below-average for the daypart.

The simple reality is that this phenomenon *can* make a big difference if the number of spots overall is small. The placement of those few "odd spots" can cause the sum of the hourly audience estimates to differ markedly from a multiple of the daypart average (in either direction). For specific stations, demos and dayparts, there can be significant variation in audience from hour to hour, so spot placement within the daypart can have a substantial impact on the schedule's estimated delivery.

Some Examples

Summarizing this phenomenon turned out to be more challenging than we had expected. Not only does the importance of this variance differ by market, daypart, demo, and station, but it also

matters tremendously how many “odd spots”⁶ there are relative to the number of hours in the daypart; the total number of spots overall also greatly influences the impact.

We struggled with a number of different ways to study and summarize this issue in a practical way. We finally settled on taking a special sample of schedules as an illustration; the more general formula follows.

For illustration purposes, we looked at a sample of:

- ▶ 3 markets (New York, Philadelphia, and Washington, D.C.)
- ▶ 5 dayparts (Morning Drive, Afternoon Drive, Mon-Fri 6AM-7PM, Weekends, and Mon-Sun 6AM-Mid)
- ▶ 5 demos (Persons 12-34, 18-34, 18-49, 35-64, and 12+)
- ▶ 20 stations sampled from the three markets

Then for each station/demo/daypart, we began by computing a number which is roughly analogous to the Relative Standard Error presented earlier. But this time, the number (a Coefficient of Variation, to be precise) describes how the hourly AQH audience estimates differ on average from the daypart estimates, expressed as a percentage of the daypart estimate. Here’s the formula for this Coefficient of Variation (CV) for Hours vs. Daypart:

$$CV = (\sqrt{(\text{Sum of } ((\text{Hrly Rtg} - \text{Dpt Rtg})^2) \div (\# \text{ Hrs.} - 1)))} \times 100 \div \text{Daypart Rating}$$

The outcome of that formula is a percentage, which is roughly equivalent to the Relative Standard Errors presented in Sections One and Two. If we assume that our measures here are normally distributed (and they appear to be sufficiently so), then one can construct confidence intervals using the values shown in earlier charts (e.g., ± 1.96 times CV equals the 95% confidence interval).

Taking an example from our work, the average Coefficient of Variation for Morning Drive for this sample was 23%. This suggests that, on average, 68% of the Morning Drive *hourly* ratings fell within $\pm 23\%$ of the Morning Drive *daypart* rating. Of course, it also means that about 32% of the hourly ratings were *more* than 23% different from the daypart rating.

Here’s how this Coefficient of Variation differed across the variables in this sample:

⁶From this point on, when we refer to “odd spots,” we’ll be assuming the situation where as many spots as absolutely possible were given even rotation. The “odd spots” are then those few leftovers—specifically, the number of spots beyond the highest possible even multiple of the number of hours in the daypart.

	Mean Observed CV Value
Persons 12-34	37%
Persons 18-34	38%
Persons 18-49	33%
Persons 35-64	33%
AM Drive	23%
PM Drive	20%
M-F 6AM-7PM	32%
M-F 6AM-Mid	49%
Weekends	52%

The mean CV overall was 35%; the median was 27%. Obviously, the breadth of the daypart can affect the CV, which reflects the amount of hour-by-hour variation within a daypart. The variations across demos are surprisingly small.

Now we have to take into account the number of spots overall, and the number of hours within the daypart. As you'll see from the table below, if there are a large number of spots, the effect of the odd spots becomes relatively small. As the number of spots grows, the effect of one or two or three spots on the total becomes smaller and smaller.

Also relevant is how many odd spots there are relative to the size of the daypart. If there's one odd spot, it can fall anywhere in the daypart, with that spot potentially deviating significantly from the daypart average. If there are two odd spots, there are fewer ways to place them evenly through the daypart, thus reducing the potential variation from the daypart average. For three spots, the average impact is even less, etc.

The table below illustrates both of those phenomena simultaneously. Because of the small variation across narrow dayparts (see above), we computed illustration data only for Morning Drive (Monday-Friday 6AM-10AM) and for the broader daypart Monday-Friday 6AM-7PM. We also present data only for the average across all demos since there was little variation on this dimension. This should present the user with a good sense of the amount of impact that hour-by-hour variation can have on the sum-of-hours calculation. (Again, we explain how to conduct this computation for other situations in a moment.)

Here's how to read the table: If only one spot is run, then its Coefficient of Variation (CV) is equal to the CV for hours compared to the daypart. In the case of Morning Drive, the placement of that one spot can cause its GRP to vary within the range of $\pm 22.6\%$ in Morning Drive 68% of

the time. For a 95% "confidence interval," multiply the CV by 1.96; in other words, 95% of the time, that one spot's hour-based GRP will be within 45% of the daypart rating.

For a larger number of spots, the reading is similar: If 5 spots are run in Morning Drive, the placement of the odd spot can cause the total hour-based GRP to differ from a daypart-based computation; 68% of the time, the hour-based GRP will be within 4.5% of the daypart-based approach, and 95% of the time, the hour-based approach will be within 8.8% of the daypart version.

Of course, where the number of spots is equal to an even multiple of the number of hours in the daypart, the CV is equal to zero. If there's equal rotation, the hourly computation method should equal the daypart approach.

Number of Spots in Schedule	Coefficient of Variation for GRPs, Hour-based vs. Daypart-Based	
	Monday-Friday 6AM-10AM	Monday-Friday 6AM-7PM
1	22.6%	31.8%
2	13.0%	21.4%
3	7.5%	16.6%
4	0.0%	13.6%
5	4.5%	11.5%
6	2.2%	9.9%
7	1.1%	8.5%
8	0.0%	7.3%
9	0.5%	6.1%
10	0.2%	5.0%
11	0.1%	3.9%
12	0.0%	2.6%
13	0.0%	0.0%
14	0.0%	2.3%
15	0.0%	1.4%

Number of Spots in Schedule	Coefficient of Variation for GRPs, Hour-based vs. Daypart-Based	
	Monday-Friday 6AM-10AM	Monday-Friday 6AM-7PM
16	0.0%	1.0%
17	0.0%	0.8%
18	0.0%	0.6%
19	0.0%	0.5%
20	0.0%	0.4%
21	0.0%	0.3%
22	0.0%	0.3%
23	0.0%	0.2%
24	0.0%	0.2%
25	0.0%	0.1%
26	0.0%	0.0%
27	0.0%	0.1%
28	0.0%	0.1%
29	0.0%	0.0%
30	0.0%	0.0%

How to Compute This Effect for Other Situations

The examples above reflect, in part, our choice of markets, dayparts, and stations. If the user needs to compute this effect for other situations, we recommend the following approach.

First, for the situation being analyzed, the user must compute the Coefficient of Variation for Hours vs. Daypart, as described above. For the particular station and daypart under study, compute:

$$CV = (\sqrt{\text{Sum of } ((\text{Hrly Rtg} - \text{Dpt Rtg})^2) \div (\# \text{ Hrs.} - 1)}) \times 100 \div \text{Daypart Rating}$$

That number provides a standardized way of describing how the hourly ratings differ from the daypart ratings for the situation under analysis. Note that this approach only captures the effect of a computational method for a particular set of data at a particular point in time.

For example, if the CV of Hours vs. Daypart is computed to be 10%, then, about two thirds of the time, any one spot would have an hourly rating that differed from the daypart average by 10% or less. In other words, the 68% confidence interval for an hourly rating's difference from a daypart rating is equal to ± 1.0 CV. A 95% confidence interval would equal ± 1.96 CVs.

Then, the user must adjust the CV for the number of total spots included in the GRP and for the number of hours in the daypart. That adjustment table is provided in Appendix A.

For example, a GRP consisting of five spots in a four-hour daypart would have a smaller amount of odd-spot error than is represented by the CV of Hours vs. Daypart. In the Appendix, the user may look up an adjustment for 5 spots in a 4-hour daypart which is equal to 20%. That means that the GRP's CV is only 20% of the CV for Hours vs. Daypart, or 2% (20% of 10%).

Conclusion

Clearly, for small numbers of spots, the placement of odd spots can cause an hour-based GRP to differ meaningfully from a GRP in which daypart audience estimates were attributed to each spot. This can be an additional and independent source of variance which users should account for in analyzing the estimated delivery of a schedule. As the number of spots grows, however, this phenomenon declines in size.

Section Four: Variance from Rounding in GRP Computation

The Problem

When we reviewed preliminary results of this study with some users, we encountered the related issue of rounding as a source of variance in the computation of GRPs. In the study reported in Section One, we used the least amount of rounding practical; our methods were generally equivalent to computing GRPs from Arbitron's new, less-rounded "client tapes," which in the near future will provide persons projections to the nearest single person. A user who wished to compute a GRP with the least possible rounding error would:

- 1) Attribute to each spot the AQH persons projection from the Arbitron tape;
- 2) Sum those projections across all spots; and,
- 3) Divide that sum by the population estimate for the demographic group.

That method would yield GRPs which had approximately the same level of precision as those we studied in Section One.

However, some users expressed a desire to know more about the additional rounding error which might be introduced by the traditional practice of simply attributing AQH *rating points* to each spot, and summing those (more rounded) estimates. Arbitron radio rating points are reported to the nearest tenth of a rating point, so the amount of rounding error in typical radio ratings is nontrivial in this application.⁷

The Theory

The relative (percentage) amount of additional variance caused by rounding error in rating-point-based GRPs is primarily a function of:

- ▶ The size of the ratings themselves;
- ▶ The number of different Arbitron ratings (from one survey period) used to compute the GRP, either different hours/dayparts or different stations; and,
- ▶ The number of spots for each unique station/daypart (which affects the relative contribution to the total GRP of each rating's rounding).

⁷There are, of course, other ways that one could compute GRPs, including the use of projections rounded to hundreds as are reported in the printed rating books; that approach would have rounding error somewhere between the GRPs we studied in Section One and the rating-point approach. But it appears that the vast majority of users who are considering regular schedule evaluation would be working with computer systems and tape-based data, in part because of the manual labor involved with the alternative. Thus, we decided to focus here on the worst-case example of rounding error likely to appear in practice.

A simple example may help to illustrate these points.

Let's start with the case of a schedule which ran entirely on one station and within one hour. In practice, all that matters for the *percentage* rounding error of the GRP is the percentage rounding error of that station's rating for that time period. If the station's rating was a 1.0, we know that it could, in fact, have ranged from a 0.95 to a 1.499... in the actual survey results. Thus, the "actual" (unrounded) rating could have been in the range "1.0 $\pm$ 5%." If 10 spots ran on that station in that hour, yielding a GRP of 10.0, the total rounding error range would still be $\pm$ 5%, since the GRP takes its rounding error from the underlying rating (only here, the *absolute* range is larger, with the GRP possibly ranging from 9.5 to 10.5).

Obviously, GRPs based on larger ratings would have less total rounding error. In the example above, if the station's rating had been a 3.0 instead of a 1.0, the rounding error would be $\pm$ 1.7% (or 3.0, plus or minus 0.05 points). If that rating were attributed to ten spots, the rounding error in the GRP total (3.0 times 10) would still be $\pm$ 1.7%.

In the examples above, the *average* rounding error in a single rating would be half of that range, or $\pm$ 2.5%; half of the original unrounded ratings would have been within 2.5% of the published rating, and half the cases would be within 2.5% to 5.0%. The incidence of 0.95 rounding to 1.0 should be equal to the incidence of 0.96 rounding to 1.0, which should be equal to the incidence of 0.97 rounding to 1.0, etc. We know of no practical reason for the distribution of unrounded data for any given rating size to be anything other than even.

Now let's consider the case of a GRP which is based on two *different* Arbitron AQH ratings. If the schedule included 5 spots in one hour with a rating of 1.0, and 5 spots in a different hour with a rating of 1.0, the GRP would be still be 10.0, but the average rounding error will be reduced. While it's still the case that either of the underlying ratings has a total rounding error of $\pm$ 5%, and the total *maximum* rounding error of the GRP is still $\pm$ 5%, we have reduced the *incidence* of the extremes because of the possibility that rounding in the two ratings will cancel out. Now, we no longer have an even distribution of possibilities; the likelihood of the "actual" (unrounded) GRP being close to 10.0 is greater. In short, the average rounding error of a two-rating GRP is less than that of a single-rating GRP, all else being equal.

The same phenomenon occurs if the user is adding together ratings for two different *stations*. The *average* percentage rounding error benefits from the inclusion of a second measure's rounding, which can sometimes offset the rounding of the first.

As more different Arbitron ratings are used in the computation of a GRP, the distribution of the rounding effect becomes more and more like a "normal," bell-like curve. A majority of rounding-error occurrences for a particular situation are clustered around the computed GRP, with less-frequent occurrences of the extreme cases. For example, if ten Arbitron ratings were used, there is *some* chance that all ten would have involved a rounding error of +0.05, but the probability of that occurrence is relatively low compared to other, smaller amounts of total rounding error.

To summarize, with a large number of different ratings used (different hours/dayparts or different stations) in the GRP, the average effect of rounding error is reduced significantly.

Examples

To help users quantify this potential source of variance for particular situations, we now provide two tools. First, we'll provide some estimates of that rounding-error variance for some simple examples. This may suffice for some users, as it describes the range of possibilities for this type of error. Then, we'll provide a formula which would allow computation of those estimates for other scenarios not shown in our example.

In both cases, we'll describe rounding-error variance in confidence-interval terms. This will allow the user to put rounding-error variance on approximately the same footing as the sampling error described earlier; namely, our approach yields a "standard error" equivalent that can be used to describe confidence intervals. That, in turn, will tell the user what plus-or-minus range captures, say, 95% of the real-world occurrences.

To do so, we'll assume that the distribution of rounding error occurrences is statistically "normal" for all but the single-rating situation (where a GRP is based entirely on one Arbitron AQH rating). In that special case, all rounding-error outcomes are equiprobable.

First, the examples.

Remember we said that the amount of rounding error in rating-point-based GRPs depends on the number of different Arbitron ratings used in the calculation (i.e., different stations or different hours/dayparts), the size of the ratings used, and the relative contribution of each rating (i.e., the number of spots to which a particular rating was applied). To provide straightforward examples in the table below, we made two major simplifying assumptions: That the rating size was the same for all stations/dayparts in a particular buy, and that the number of spots was equal across stations and across hours/dayparts. As noted earlier, we'll shortly provide a formula for estimating other situations.

So, in the table which ends this section, we present 95% confidence interval estimates for the rounding-error variance of a number of simple buys:

- ▶ Seven different numbers of hours/dayparts in the GRP computation;
- ▶ Seven different sizes of AQH rating (again, assuming all of the ratings used were of the same size); and,
- ▶ Three different numbers of stations contributing to the GRP.

Because of our simplifying assumptions, the number of *spots* per sé is not shown on the table. What matters is the number of different Arbitron AQH ratings used in the calculation, which is captured in the column headings (for hours/dayparts), and in the number-of-station rows.

Here's how to read an example from the table. If a GRP was calculated from...

- ▶ four different hours/dayparts (the column labeled "4"), and
- ▶ two different stations (one of the rows labeled "2 stations"), for a total of 8 different Arbitron ratings contributing to the GRP, and

- ▶ each Arbitron AQH rating equaled 1.0 (which would lead you to the eleventh row of data),

...then the GRP computed from those numbers would fall within 2.0% of the unrounded GRP total 95% of the time. Conversely, 5% of the time, the rounding error could cause the GRP unrounded total to be more than 2.0% different from the rounded total.

That source of variance would be separate and apart from the other sources of variance described in previous sections.

The Formula

To actually compute a confidence interval for rounding-induced variance, the user must know the following, in addition to the GRP itself:

S	=	Number of spots run on each station
H	=	Number of different hour/daypart ratings used in the computation for each station

Then, for cases involving more than one station and/or more than one hour/daypart rating:

$$\begin{aligned}
 SE_{\text{rounding}} &= \text{Absolute Standard Error of rounding variation for rating-based GRPs} \\
 &= 0.029 \times \sqrt{(\text{Sum of } (S^2 \div H \text{ for each station}))}
 \end{aligned}$$

Where:

0.029	=	Standard deviation of Arbitron rating rounding error, assuming a normal distribution of rounding error.
-------	---	---

The derivation of the formula is available upon request.

To compute a confidence interval:

±1.00 Standard Error	=	68% confidence interval
±1.65 Standard Errors	=	90% confidence interval
±1.96 Standard Errors	=	95% confidence interval
±2.58 Standard Errors	=	99% confidence interval

This would express the confidence interval in absolute points. A relative (percentage) value could be obtained by dividing the confidence interval values by the GRP value and multiplying by 100.

Using the example cited earlier (four different hours/dayparts and two stations), we would insert the values into the formula as follows, using arbitrary equal values for number of spots (S=8):

$$\begin{aligned} SE_{\text{rounding}} &= 0.029 \times \sqrt{(\text{Sum of } (S^2 \div H \text{ for each station}))} \\ &= 0.029 \times \sqrt{((8^2 \div 4) + (8^2 \div 4))} \\ &= 0.029 \times \sqrt{(32)} \\ &= 0.164 \end{aligned}$$

The 95% confidence interval equals $\pm 1.96 \times 0.164$, or ± 0.321 rating points. Since we ran 8 spots per station, each of which has a rating of 1.0, we have a GRP here of 16.0, so the 95% confidence interval expressed as a percentage is $\pm 2.0\%$, the number which appears on the table on the next page.

Conclusions Concerning Rounding Error

Clearly, rounding error can introduce an additional source of variance to a radio-schedule analysis if one computes GRPs from Arbitron's AQH rating estimates. Arbitron radio ratings are reported to the nearest tenth of a rating point, and that involves more underlying rounding than the relatively unrounded persons projections soon to be available on Arbitron client tapes.

If a user finds it necessary to compute GRPs by adding rating points, then the analysis above should be taken into account. That rounding error constitutes an additional, independent source of variance that could cause a GRP to differ from expectations in ways that can now be quantified.

**Examples of 95% Confidence Intervals for
Rounding-Error Variance of Rating-Point-Based GRPs**

(Values Shown are Plus-or-Minus Percents Describing 95% of Rounding-Error Occurrences)

Size of Rating	No. of Diff. Stations	No. of Different Hours/Dayparts in Calculation						
		1	2	3	4	6	12	18
0.1	1	47.5%	40.2%	32.8%	28.4%	23.2%	16.4%	13.4%
	2	40.2%	28.4%	23.2%	20.1%	16.4%	11.6%	9.5%
	5	25.4%	18.0%	14.7%	12.7%	10.4%	7.3%	6.0%
0.2	1	23.8%	20.1%	16.4%	14.2%	11.6%	8.2%	6.7%
	2	20.1%	14.2%	11.6%	10.0%	8.2%	5.8%	4.7%
	5	12.7%	9.0%	7.3%	6.4%	5.2%	3.7%	3.0%
0.5	1	9.5%	8.0%	6.6%	5.7%	4.6%	3.3%	2.7%
	2	8.0%	5.7%	4.6%	4.0%	3.3%	2.3%	1.9%
	5	5.1%	3.6%	2.9%	2.5%	2.1%	1.5%	1.2%
1.0	1	4.8%	4.0%	3.3%	2.8%	2.3%	1.6%	1.3%
	2	4.0%	2.8%	2.3%	2.0%	1.6%	1.2%	0.9%
	5	2.5%	1.8%	1.5%	1.3%	1.0%	0.7%	0.6%
2.0	1	2.4%	2.0%	1.6%	1.4%	1.2%	0.8%	0.7%
	2	2.0%	1.4%	1.2%	1.0%	0.8%	0.6%	0.5%
	5	1.3%	0.9%	0.7%	0.6%	0.5%	0.4%	0.3%
5.0	1	1.0%	0.8%	0.7%	0.6%	0.5%	0.3%	0.3%
	2	0.8%	0.6%	0.5%	0.4%	0.3%	0.2%	0.2%
	5	0.5%	0.4%	0.3%	0.3%	0.2%	0.1%	0.1%
10.0	1	0.5%	0.4%	0.3%	0.3%	0.2%	0.2%	0.1%
	2	0.4%	0.3%	0.2%	0.2%	0.2%	0.1%	0.1%
	5	0.3%	0.2%	0.1%	0.1%	0.1%	0.1%	0.1%

Section Five: Technical Notes, Caveats and Other Miscellany

Averaging Books for More Reliability

A number of users have asked about the procedures to follow if the user wishes to average two surveys together for increased reliability. Fortunately, our model is easily extensible to that application.

There's really no difference in approach, only in the values to be input into the models. In the formulas shown in Section One, one simply inputs the two-survey average measures (GRP, Reach, or Frequency), along with the two-survey *total* of the in-tab for that demographic and market. All the other values, including the Statistical Efficiency from the look-up table, remain the same.

A purist might argue that if the in-tab samples of the two surveys differed significantly in size (say, as a result of Arbitron sample size increases), one should account for the additional "weighting" involved in the averaging process. But we doubt that there's much to be gained from that extra precision.

Two-Tail vs. One-Tail Probabilities

In all of the work presented on confidence intervals, we opted for the use of a "two-tail" (nondirectional) definition of a confidence interval or of significance. Since the research hypothesis involves potential sampling error in either direction, the two-tail approach seems appropriate. However, some users might wish to apply one-tail tests of differences if their hypothesis is only in one direction. That would result in a somewhat greater likelihood of concluding that "real" change had occurred (vs. only sampling error). We will be glad to provide the appropriate confidence-interval multiples to those users on request; they're also available in most common statistical texts and references.

Market-Level Variation in Efficiency

To make the application of our Section One and Section Two models practical, we chose to ignore the fact that Statistical Efficiencies can vary slightly by market and by survey period. Those variations can be caused in part by variations in the amount of sample balancing and in part by the amount of actual variation in listening within the market at that point in time. Unlike *Replication II* and its application to Arbitron's regular syndicated market reports, we have not attempted to provide survey-specific market-level adjustment factors to our model. The data necessary for that adjustment are not routinely available to the likely users of this report.

What small market-by-market differences in Efficiency we did see in the current study did not correlate with any obvious Metro-level factors (e.g., market size), and would be partly accounted

for by other factors in our model (demographics, rating size, etc.). Furthermore, the *average* amount of sample balancing is accounted for in the Statistical Efficiencies provided. Overall, Dr. Soong felt that we could comfortably ignore Metro variation per sé.

Reach vs. Cume

Astute readers may notice that we *did* decide to account for the size of a daypart in our Reach model, unlike *Replication II* which concluded that daypart is an unimportant variable for cume reliability. Our model reflects the fact that Reach isn't quite a cume estimate, as discussed further in Section One.

List of Markets

The database for the study reported in Section One consisted of the syndicated diaries from the Winter 1995 Arbitron survey in 30 markets selected from the top 50 (largest) Metros. The markets used were:

New York	Houston-Galveston
Los Angeles	Denver-Boulder
Chicago	Seattle-Tacoma
Philadelphia	Kansas City
San Francisco	Milwaukee-Racine
Detroit	Atlanta
Boston	Portland, OR
Washington, DC	Phoenix
St. Louis	San Diego
Cleveland	Sacramento
Baltimore	Memphis
Pittsburgh	Providence-Warwick-Pawtucket
Dallas-Ft. Worth	Tampa-St. Petersburg-Clearwater
Minneapolis-St. Paul	Riverside-San Bernardino
Cincinnati	Miami-Ft. Lauderdale-Hollywood

Applicability to Markets Ranked 50+

Because of the relative lack of variation in Statistical Efficiency by market, we are confident that the results of this Top-50-based study are generalizable to smaller markets. The most important differences in smaller markets—rating size and sample size—are accounted for in the formulas provided.

It's possible, of course, that slightly different Statistical Efficiency values for the model would have resulted if we had chosen different markets for our study. But we believe the study's large scale makes the likelihood of *meaningful* variations very, very small.

Acknowledgments

Obviously, we are grateful to Dr. Roland Soong for his quick and imaginative work on this project. The words in this report reflect the views of Arbitron, but the fundamental statistical thinking and the immense computational efforts required for the Section One analysis are largely a result of Dr. Soong's special talents.

In addition, we should acknowledge the in-house statistical contributions of Arbitron's Senior Statistician, Thomas White. Mr. White has been invaluable to the success of this project, as he contributed solutions to a number of thorny problems. Among other things, he was responsible for the formula which allows us to provide Standard Errors for Frequency, and for the analyses of the effect of "odd spots" and of rounding. Mr. White also provided the interpolations that allowed us to extend the work done by Dr. Soong to additional number-of-quarter-hour sets of Statistical Efficiencies, and he developed the general-case formula that accounts for number-of-stations efficiency effects without assumptions of even numbers of spots.

§

We hope that users will find this report useful. If you have any questions or any other comments after reading this report, please don't hesitate to communicate them directly to:

James D. Peacock
Vice President, Research
The Arbitron Company
9705 Patuxent Woods Drive
Columbia, MD 21046-1572

(410) 312-8433 (voice)
(410) 312-8604 or -8624 (fax)

Appendix A: Adjustment Factors for Odd-Spot Variation

In "Section Three: Variance from Comparing Hour-Based GRPs with Daypart Expectations," we described how to compute a Coefficient of Variation for Hours vs. Dayparts. That CV must be further adjusted before applying to a GRP by taking into account the number of hours in the daypart, and the number of spots within the daypart, using the reduction factors below. This table takes into account how many "odd spots" are *possible* with a given combination, and how much *influence* they could have on a particular schedule.

No. of Spots in Daypart	Number of Hours in the Daypart				
	4	5	8	13	18
1	100%	100%	100%	100%	100%
2	58%	62%	66%	67%	68%
3	33%	41%	49%	52%	54%
4	0%	25%	38%	43%	45%
5	20%	0%	29%	36%	39%
6	19%	17%	21%	31%	34%
7	14%	18%	14%	27%	30%
8	0%	15%	0%	23%	27%
9	11%	11%	11%	19%	24%
10	12%	0%	13%	16%	22%
11	9%	9%	13%	12%	19%
12	0%	10%	13%	8%	17%
13	8%	9%	11%	0%	15%
14	8%	7%	9%	7%	13%
15	7%	0%	7%	9%	11%
16	0%	6%	0%	10%	9%
17	6%	7%	6%	10%	6%

No. of Spots In Daypart	Number of Hours in the Daypart				
	4	5	8	13	18
18	6%	7%	7%	10%	0%
19	5%	5%	8%	10%	5%
20	0%	0%	8%	9%	7%
21	5%	5%	7%	9%	8%
22	5%	6%	6%	8%	8%
23	4%	5%	4%	7%	8%
24	0%	4%	0%	6%	8%
25	4%	0%	4%	4%	8%
26	4%	4%	5%	0%	8%
27	4%	5%	5%	4%	8%
28	0%	4%	5%	5%	8%
29	3%	3%	5%	5%	7%
30	4%	0%	4%	6%	7%
31	3%	3%	3%	6%	6%
32	0%	4%	0%	6%	6%
33	3%	4%	3%	6%	5%
34	3%	3%	4%	5%	4%
35	3%	0%	4%	5%	3%
36	0%	3%	4%	4%	0%
37	3%	3%	4%	4%	3%
38	3%	3%	3%	3%	4%
39	3%	3%	3%	0%	4%
40	0%	0%	0%	3%	4%
41	2%	2%	2%	3%	5%

No. of Spots in Daypart	Number of Hours in the Daypart				
	4	5	8	13	18
42	3%	3%	3%	4%	5%
43	2%	3%	3%	4%	5%
44	0%	2%	3%	4%	5%
45	2%	0%	3%	4%	5%
46	3%	2%	3%	4%	5%
47	2%	3%	2%	4%	5%
48	0%	3%	0%	4%	4%
49	2%	2%	2%	3%	4%
50	2%	0%	3%	3%	4%

Appendix B: Restrictions and Other Limitations

General Information and Limitations/ The *Arbitron Study of Radio-Schedule Audience Estimate Reliability* is provided by Arbitron to Arbitron clients and is intended to provide an aid in determining estimated reliability of aggregated audience estimates of radio-schedule audience delivery. This study was conducted by Dr. Roland Soong of Audits & Surveys Worldwide, under contract with Arbitron, with additional analyses provided by Arbitron's Research Group. The findings of this study provided herein are based on Arbitron Winter 1995 Radio survey respondent data and listening information from 104,166 total in-tab diaries from 30 of the top 50 Arbitron Radio Metros as processed for Arbitron Winter 1995 Radio Market Reports.

The Arbitron data, information and audience estimates used for this study are subject to the statistical variances associated with surveys which use a sample of the universe and, additionally, to all of the factors described on Page 5B and/or the *Limitations* section on Page iii of the applicable Winter 1995 Radio Market Reports. Because this study is based on a single survey of 30 selected Arbitron Radio Metros, the study results provided herein may differ if based on a different set or subset of Arbitron Radio Metros and/or surveys.

Users of this study should become familiar with the *Description of Methodology* and *Limitations* sections printed on Pages i-iv of Arbitron Radio Market Reports for the Winter 1995 survey which are applicable to the in-tab diary sample on which this study is based. Additional details on Arbitron methodology may also be found in a separate publication, available to all syndicated radio report subscribers, titled *Description of Methodology for Radio*.

Users of this study should also note that all audience estimates and their statistical evaluators, including reliability estimates, are approximations subject to statistical variations and other limitations. Audience estimates and their estimated reliability cannot be determined to any precise mathematical value or definition.

Warning: All Arbitron Audience Estimates Are Copyrighted and Proprietary/ Each Arbitron audience estimate and reliability estimate provided herein is copyrighted by and proprietary to The Arbitron Company. The unauthorized use of any Arbitron audience estimate or any reliability estimate provided by this study constitutes copyright infringement which could subject the infringer to statutory damages of up to \$100,000 and criminal penalties of up to one year imprisonment and a \$25,000 fine pursuant to Chapter 5, Sections 504 and 506 of Title 17 of the U.S. Code. All users of this study are referred to the Restrictions on Use of Report section below.

Disclaimer of Warranties/ Arbitron makes no warranties, express or implied, including without limitation any warranty of merchantability or fitness, concerning: data gathered or obtained by Arbitron from any source; the present or future methodology employed by Arbitron in producing Arbitron audience estimates and/or reliability estimates of Arbitron audience estimates; or the Arbitron data, audience estimates or reliability estimates contained herein. All Arbitron data, audience estimates and reliability estimates contained herein represent only the opinion of Arbitron and reliance thereon and use thereof shall be at the user's own risk.

Limitation on Liability/ The sole and exclusive remedy for Arbitron's liability of any kind, including without limitation liability for any warranty of merchantability or fitness or for negligence with respect to this study shall be limited to an amount not to exceed \$500. In no event shall Arbitron be liable for incidental or consequential damages or injunctive relief.

Restrictions on Use of Report/ All Arbitron data, audience estimates and reliability estimates are copyrighted by and proprietary to Arbitron. Arbitron data, audience estimates and reliability estimates may only be used by Arbitron clients who are also subscribers to the Arbitron Radio Market Report and pursuant to the restrictions and limitations on use printed herein and printed in Arbitron Radio Market Report for Winter 1995 unless superseded by Arbitron Radio Market Reports published thereafter. All Arbitron data, audience estimates and reliability estimates are for the exclusive use of licensed users of the Arbitron Radio Market Report and their representatives and may be disclosed only to advertisers, prospective advertisers and their agencies for the purpose of obtaining and retaining advertiser accounts and through advertising or promotional literature. Any commercial use of Arbitron data, audience estimates or reliability estimates for the purpose of selling advertising time or space by or on behalf of broadcast, cable or print media must be under the terms of a written license agreement between that medium and Arbitron specifying permitted uses. For an Arbitron client to divulge any Arbitron data, audience estimates or reliability estimates to a nonsubscribing client, or to lend and/or give a copy and/or a reproduction of any part of this study to any nonsubscriber, including print media, advertisers and/or their agencies constitutes a violation of copyright law. Quotations of Arbitron data, audience estimates and reliability estimates as permitted hereunder for purposes of advertising or promotion must identify Arbitron as the source and that the Arbitron data, audience estimates and reliability estimates are copyrighted. Users of Arbitron data, audience estimates and reliability estimates should also mention that these data, audience estimates and reliability estimates are subject to all qualifications and limitations stated herein and in the Arbitron Radio Market Report for Winter 1995 unless superseded by Arbitron Radio Market Reports published thereafter. Arbitron data, audience estimates and reliability estimates may not be used in any manner by nonclients of Arbitron without written permission from Arbitron. Users of Arbitron data, audience estimates and reliability estimates are referred to the current policies of the Federal Trade Commission relating to the use of audience estimates.

ARBITRON

New York

142 West 57th Street
New York, New York 10019-3300
(212) 887-1300

Chicago

211 East Ontario, Suite 1400
Chicago, Illinois 60611-3289
(312) 266-4150

Atlanta

9000 Central Parkway, Suite 300
Atlanta, Georgia 30328-1639
(770) 551-1400

Los Angeles

5670 Wilshire Blvd., Suite 600
Los Angeles, California 90036-5606
(213) 932-6500

Dallas

One Galleria Tower
13355 Noel Road, Suite 1120
Dallas, Texas 75240-6646
(214) 385-5388

Washington/Baltimore

9705 Patuxent Woods Drive
Columbia, Maryland 21046-1572
(410) 312-8000

A division of Cerdian Corporation

© 1995 The Arbitron Company



Printed in the USA
on recycled paper

AD-1256 2.5M 10/95