Arbitron Replication: A Study of the Reliability of Broadcast Ratings

Arbitron™ Replication: A Study of the Reliability of Broadcast Ratings

American Research Bureau, Inc.
New York
1974

Copyright © 1974 ARB (American Research Bureau), Inc.
All rights reserved
including the right of reproduction
in whole or in part in any form.
Published by American Research Bureau, Inc.
1350 Avenue of the Americas
New York, N.Y. 10019

Library of Congress Catalog Card Number: 74-83260
Design by Ong & Associates, Inc.
Manufactured in the United States of America
Typesetting by Towne Typographers, Inc.
Printed by Lerman Graphics, Inc.
First Edition

Preface

Preface

Two of the most fundamental areas of concern to the user and producer of Arbitron Television audience estimates are the kinds of data that are produced and the reliability of these data. As researchers, both worry about sampling, weighting, processing, reporting—all of the areas which affect the reliability of audience estimates—as well as the kinds of estimates produced.

In some quarters, much time is given to discussions of the kinds of estimates to be produced, with only limited consideration given to the reliability of the estimates. As a result, many new kinds of audience estimate breaks have been proposed and/or produced, with less than a complete understanding of the reliability of these estimates.

It was not too long ago that audience estimates were produced only for total households, people, men, women, and children. But the demand grew for additional and more descriptive demographics by sex and age definitions, everyone knowing full well that the sample size available for these kinds of estimates got smaller as the definition of a sex/age group became more restricted.

The problem of determining the reliability of these estimates was complicated by the fact that only rough approximations of the effective sample size or base (ESB) for these demographic groups could be calculated. Thus, sampling error, or Standard Error, measures for these estimates were very imprecise.

In the late sixties, the Broadcast Rating Council (BRC) began discussing with Arbitron and other groups, the concept, meaning and calculation techniques of ESB's. Although this was the initial approach to reliability, the overall complexity of the ESB concept precluded significant progress.

In 1971, the National Association of Broadcasters (NAB), joined by the BRC, challenged the rating services to study the reliability of their published estimates.

The Broadcast Rating Council had become particularly interested in the reliability concept as it pertains to the various demographic data columns that appear in the published audience measurement books. The BRC was concerned that, in recent years, the number of demographic columns had continued to increase, but very little had been done to increase the sample size upon which these new data columns were developed. The BRC used the reliability concept as a means to question whether certain small demographic data columns were precise, and not misleading to unwary users.

Arbitron accepted this combined challenge and began to investigate ways to empirically measure the reliability of its audience estimates, and in the process, develop techniques for determining more precisely the ESB for any demographic group reported upon in the Arbitron Market Reports. In conjunction with MarketMath, Inc., we found that the best way to determine the reliability of Arbitron audience estimates was through the replicated subsamples, or replication, procedure.

Arbitron's replication study was designed to investigate the reliability of every demographic audience data column in our published Market Reports. In addition, the study was designed to accommodate each kind of estimate published by Arbitron.

It is only now, after three years of work by Arbitron and MarketMath, involving many months of computer time and statistical analysis, and many thousands of dollars, that a truer measure of the reliability of all Arbitron Television audience estimates, including demographic data and the various data types, becomes available.

The replication analysis has enabled us to measure empirically the statistical efficiency of our data, which thereby enables us to compute precisely ESB's for all demographic categories. This was possible through the discovery of the factors which are significant in the determination of the reliability of Arbitron audience estimates.

The study and analyses completed to date form a starting point for communicating the results of Arbitron's investigation of the reliability of its Television audience estimates to the broadcast advertising industry.

Please note that the analyses and findings from our investigations to date have been based on a national sample of television households. We in no way intend to imply that the formulas and factors developed from the national sample are completely applicable to each and every individual television market. We used the national sample as a starting point, not as the final point, in the investigations.

At this writing, Arbitron is conducting replication analyses for a group of individual television markets (selected by the

BRC), which will be used to test the applicability of the national sample results to local market situations. These analyses may result in the identification of variables affected by local market conditions.

Replication analyses will not be completed for all individual television markets; this would be impractical. Consequently, a modelling system must be developed in order to provide accurate estimates of ESB's and Standard Errors applicable to data published at each local market level. Our national sample analysis, followed by individual market testing and modelling, is intended to provide such a model.

The Arbitron investigations have studied cume ratings and average ratings. No new relationships were discovered in the reliability of cumes. Significant concepts, which previously had been thought to be true, have now been empirically demonstrated in the study of the reliability of average ratings. Our comments in the report which follows, therefore, are applicable *ONLY* to average ratings and other averaged estimates, and their increased reliability as an evaluative tool.

What can be considered the ultimate practical value of a technical investigation of the reliability of Television audience estimates to the user of audience research data?

Using samples to measure any kind of behavior always involves statistical errors and biases. Arbitron's goal is to keep these errors and biases at such low levels as to minimize the total error surrounding the audience data we produce. This is to ensure that Arbitron audience data are as reliable as possible. If the data are more reliable, there will tend to be less variation in data between surveys that is the result of sampling and related errors and biases.

If the variability of data due to sampling error is smaller than expected, then what the data tell the user becomes much more important. For example, if a given rating is steady and can be shown to have high reliability (low Standard Error), then the user of the data—both buyer and seller—can be much more confident in what he is buying or selling. If he is more confident, it is possible for the latitude given in making a buy or a sell to be changed to restrict the degree of audience fluctuation allowable between surveys. Also, since ratings data are the criteria for measuring the success of a particular programming effort, if they are found to be more reliable, then the station can evaluate new programming concepts with a greater degree of confidence.

Considering the situation where a very reliable piece of data consistently varies by a large amount from survey to survey, this situation should indicate to the buyer and the seller that something important is happening to that particular rating group. And that whatever the cause, the change in audience is real! This could certainly cause the re-evaluation of several operating premises that prevail within the industry.

Although Television estimates are the primary point of discussion in this report, Radio estimates have not been forgotten. We have also studied the reliability of Arbitron Radio audience estimates through the replication procedure. These studies have not proceeded as far as the Television studies, but we have completed sufficient analyses to make note of the results found to date in the report. Most of the work and essential findings for the Radio studies are identical or similar to those for Television.

Arbitron does not consider this replication work to be the "last word" in the measure of its audience data reliability. We do believe, however, that this work represents a significant contribution to the study of Arbitron broadcast audience data reliability. This is a further step in our desire to define broadcast audiences in such a way that advertisers and stations can have greater confidence in the kinds of data that are produced.

Finally, we must recognize and thank Jerome Greene and the staff of MarketMath, Inc., for Appendices B, C, and D to this report and for their continued perseverance and patience with us during the development and analysis of this approach. A special note of thanks and appreciation goes to Dr. Martin Frankel and Russ McKennan for their diligence to perfection which guided some of our investigations. To all who counselled us and asked critical questions, we say thank you and hope this report reflects such contributions.

American Research Bureau
Beltsville, Maryland
April 1974

R. D. Altizer
R. R. Ridgeway, Jr.

Table of Contents

-

Table of Contents

Arbitron Replication: A Study of the Reliability of Broadcast Ratings

Introduction

Introduction

Over the past three years, the American Research Bureau has been investigating one of the most fundamental questions in syndicated television audience research:

How well does a sample of respondents represent the viewing behavior of the total population from which the sample is drawn?

This question relates to the reliability, or precision, of sample measurement of television audiences, whether we are talking about . . .

- actual audience estimates or levels;
- changes in audience estimates from one survey to another; or
- differences between station audiences based on the same survey.

Our investigations began with a pilot study of the reliability of Arbitron Television audience estimates using a sample from an individual market. With the experience gained from this pilot study, we investigated the reliability of Arbitron Television audience estimates using a national sample of television households.

The national sample was used as a starting point in our investigations, because we felt the resulting data would be more applicable and generalizable to all individual markets than a sample of only a few specific markets. The national sample enabled each market to play a contributory role in the final results. Thus, averages developed from this sample should be more reflective of the national picture than averages developed from selected individual markets.

In this report, we describe and discuss . . .

- the procedures followed in determining the reliability of Arbitron Television audience estimates using the national sample;
- the analysis of the resulting data;
- the conclusions drawn from this analysis;
- the major implications from this study for the data user; and
- the procedure for implementing the results of the study to determine more accurately the reliability of any published Arbitron Television audience estimate.

The report is organized into five basic parts, as described below:

Chapter I —a general summary of the study.

Chapter II —a comprehensive report on the study, describing in detail the procedures used and analyses completed.

Chapter III—a discussion of how the results of the study can be implemented by the user of Arbitron Television audience estimates.

Chapter IV—a discussion of our investigations of the reliability of Arbitron Radio audience estimates completed to date and the implications of the results for Arbitron Radio data users.

Chapter V —a presentation of further details of the methodological and statistical procedures utilized in the study.

Throughout this report, we refer to the terms "Arbitron Television audience estimates", "audience estimates", and "estimates". By these terms, we mean the common measures of television audience size as described below for your reference.

Households Using Television (HUT)—the percent of unduplicated households (with one or more sets tuned in) which viewed all television stations combined for five or more minutes during the average quarter-hour of the time period involved. HUT is expressed as a percentage of the total number of television households in the reported survey area.

Rating—the percent of television households or persons in a particular sex/age category viewing a station for five or more minutes during an average quarter-hour of the time period involved. The rating is expressed as a percentage of the total number of television households or persons in the sex/age category in the reported survey area.

Station Share—the percentage of the total Households Using Television (HUT) reached by a station during the specified time period.

Projection—the estimated number (in thousands) of households or persons in a particular sex/age category viewing a station, or all stations combined, for five or more minutes during an average quarter-hour of the time period involved.

Chapter I

Summary and Conclusions—
Major Implications for the User

Summary and Conclusions—
Major Implications for the User

A study of the reliability of television audience estimates is a study of how well a sample of respondents represents the viewing behavior that would have resulted if a census had been conducted in the same manner and with the same care as the sample.

The reliability, or precision, of television audience estimates is expressed in terms of *sampling error*—the plus-minus limits within which we can be confident that the estimate represents the total population on which it is based.

At present, the amount of sampling error involved in television survey estimates is measured by a statistical formula which assumes that the estimates are derived from an unweighted simple random sample—which is not the true situation involved in the survey. Thus, our current measures of sampling error are at best rough *approximations* of the reliability of reported television audience estimates.

Arbitron Television wanted to determine more accurately the amount of sampling error involved in its published audience estimates. To accomplish this task, MarketMath, Inc., Arbitron's statistical consultants, was commissioned. Through MarketMath's investigations, we determined that the best way to investigate the true reliability of television audience estimates was through a procedure referred to as *replication.*

Replication involves dividing a total in-tab sample into mutually exclusive random parts, or subsamples, processing television audience estimates for each subsample, and comparing statistically the resulting estimates.

The replication procedure was carried out using a national sample of in-tab diaries from the February/March 1972 Arbitron Television nationwide survey.

Resulting from the replication procedure were thousands of numbers called *Statistical Efficiencies.* These numbers express the relationship between the estimated amount of sampling error, calculated using a theoretical or hypothetical (simple random sample) formula, and the true amount of sampling error, calculated using data from the national sample for each of the replicated subsamples.

Through the Statistical Efficiency value, which is based on actual empirical data, we can determine more accurately the amount of sampling error around a television audience estimate.

The Statistical Efficiency value is multiplied times the actual in-tab sample size to compute the true *Effective Sample Base* (ESB) of the survey sample—that is, determine the sample size (simple random sample) the actual in-tab sample is performing as. If we had a Statistical Efficiency of 1.0, the ESB would equal the in-tab sample size. If the Statistical Efficiency value were 4.0, the ESB would equal four times the in-tab sample size; in other words, the sample would be performing as if it were four times as large as it actually is.

For a given audience estimate, the amount of sampling error *decreases* as the size of the ESB *increases.* So, with Statistical Efficiencies greater than 1.0, the sampling error of an estimate becomes smaller, and its reliability thus becomes greater.

What our study of the reliability of Arbitron Television audience estimates has shown is that average ratings, and other averaged estimates, are more precise than indicated by current approximation procedures.

We have found through replication analysis that, except for cume estimates[1] and estimates based on an average of up to three quarter-hour observations, Statistical Efficiency values of Arbitron Television audience estimates are generally greater than 1.0, showing that Effective Sample Bases are greater than in-tab sample sizes. This means that Arbitron Television survey samples are performing as if they were larger than the simple total of respondents in the sample.

Because they are more precise, we can be more confident that these audience estimates truly reflect the viewing behavior of the total population and that changes in reported audience sizes from survey-to-survey are a function of changes in actual viewing behavior, and not sampling error.

With this knowledge, our goal was to develop a method of determining *accurately* Statistical Efficiency values for *all* Arbitron Television audience estimates without having to repeat the replication procedure. Following an extensive analysis of the results of the replication procedure, we were able to develop a mathematical model which does just that. Through the model, we can determine precisely the Statistical Efficiency value of an estimate given two bits of information:

[1] As used here, cume estimates refer to those based on a one quarter-hour observation.

(1) **The number of quarter-hours averaged to compute the audience estimate.**

As the number of quarter-hours averaged to develop an estimate *increases,* the Statistical Efficiency *increases.*

Audience estimates for one quarter-hour are based upon a single observation or sample of the total respondent sample. Estimates for more than one quarter-hour are based upon more than one observation of the total respondent sample. The more observations made before the data are combined, the more stable the average will be—and in turn, the higher the Statistical Efficiency will be.

(2) **The population group upon which the audience estimate is based.**

Smaller, more tightly-defined demographic groups tend to have higher Statistical Efficiencies than larger, less tightly-defined demographic groups.

This is because we are more efficient in a statistical sense in measuring the viewing behavior of smaller, more tightly-defined demographic groups.

The tighter the demographic definition of the population group, the less likely we are to find two or more people in this group who live in the same household. Thus, there is less "clustering" effect in repeated observations of viewing in the same household.

In addition, smaller and more tightly-defined demographic groups are somewhat more likely to view television during the same time period. Thus, there is more efficiency in measuring these groups' viewing behavior through survey sampling.

We have applied this knowledge to develop a table of Statistical Efficiency values which can be used with any audience estimate reported in Arbitron Television Market Reports to determine more accurately the sampling error around the estimate. This table is presented in Chapter III of this report, along with a description of the procedure to determine the sampling error around individual audience estimates.

Implications

As a result of the investigation of the reliability of Arbitron Television audience estimates, and the discovery of the key interacting variables of this reliability, we have determined that certain kinds of Arbitron Television audience estimates are much more reliable than current approximation procedures would lead us to believe.

Users of Arbitron research data can now have a clearer understanding of the true reliability—and source of variability—of Arbitron Television average audience estimates. They can therefore make better evaluations of the data published in Arbitron Television Market Reports and have more confidence in decisions based upon these data.

Users can now also spot trends in television audiences more rapidly and be more confident that the data in Arbitron Television Market Reports reflect the audiences which received the commercial messages delivered.

Key to these implications is the expanded use of average estimates as an evaluative tool. Average estimates which are stable or have a definite trend can be used with greater confidence earlier in the decision-making process.

The conclusions discussed for Arbitron Television audience estimates appear to be just as applicable for Arbitron Radio audience estimates. The implications for Radio, however, could be even greater. A considerable number of published Radio audience estimates are based on relatively small population groups (and thus small sample sizes). But because most of the published estimates for these groups are based on a large number of averaged quarter-hours (and therefore have relatively high Statistical Efficiencies), the estimates are much more reliable than current approximation procedures would indicate.

Chapter II

A Comprehensive Discussion of the Study

A Comprehensive Discussion of the Study

A. An Overview of the Study

A study of the reliability of television audience estimates is a study of the precision with which measurement of the viewing behavior of a sample of respondents represents the total population from which the sample is drawn.

This reliability or precision is expressed in terms of sampling error—the plus-minus limits within which we can be confident of the audience estimate. Mathematically, we determine the size of the sampling error by a statistic referred to as the *Standard Error.*

At present, Standard Errors of television audience estimates are calculated using a formula which assumes that the estimates are derived from an unweighted simple random sample—which is not the true situation involved in the survey. For this reason, the resulting Standard Error is at best a rough *approximation* of the reliability of the rating.

The goal of the present study was to develop a procedure to determine with more accuracy Standard Errors of Arbitron Television audience estimates and ultimately use this procedure to provide the user with the information he needs to better assess the reliability of Arbitron Television audience estimates.

To accomplish this task, Arbitron commissioned its statistical consultants, MarketMath, Inc. Through MarketMath, we found that our goal could best be met by studying Standard Errors empirically developed through the *replicated subsamples,* or *replication,* procedure. These empirical Standard

Errors could best be developed, summarized, and evaluated in terms of a more fundamental measure, that of *Statistical Efficiency* values of Arbitron Television audience estimates.

The Statistical Efficiency value expresses the relationship of the estimated sampling error of an audience estimate calculated from a theoretical or hypothetical (simple random sample) formula and the true sampling error calculated from empirical data (i.e., actual Arbitron Television audience estimates).

The replication procedure was carried out using a national sample of in-tab diaries from the February/March 1972 Arbitron Television nationwide survey.

Upon completion of the replication portion of the study, the thousands of resulting Statistical Efficiencies were extensively analyzed to determine how their utility could be maximized for Arbitron Television audience data users. These analyses culminated in the development of a general mathematical model which is capable of predicting precisely Statistical Efficiency values of Arbitron Television audience estimates, based upon the variables:

(1) the number of quarter-hours averaged to compute the estimate; and

(2) the population group upon which the estimate is based.

This model was applied to calculate a table of Statistical Efficiency values which can be used to compute the true Standard Error, or reliability, of any audience estimate published in current Arbitron Television Market Reports, knowing only the two determinants discussed in the previous paragraph.

In the next two sections, a detailed report on the concepts and procedures involved in the study of the reliability of Arbitron Television audience estimates is presented.

In the first section, we discuss the major concepts of the study and their meaning. Here we deal with:

(1) Sampling Error and Effective Sample Bases, and their application in broadcast audience research; and

(2) Statistical Efficiency and its key role in the determination of more accurate Standard Errors of Arbitron Television audience estimates.

In the second section, we discuss the procedures used in the calculation and analysis of the data for the study. Here we deal with:

(1) The procedure by which Statistical Efficiencies were developed using actual empirical data;

(2) The analysis of the resulting Statistical Efficiencies for application in syndicated Arbitron Television surveys; and

(3) The general mathematical model from which a table of Statistical Efficiency values for determining more accurately the Standard Error of any Arbitron Television audience estimate was developed.

Following these discussions, you will find the last three major parts of this report. In Chapter III, we describe how the results of this study can be implemented by the user of Arbitron Television audience estimates to calculate a more accurate Standard Error of any Arbitron Television audience estimate and to apply the numbers used to calculate the Standard Error of individual estimates to determine if the difference between two estimates is statistically significant.

In Chapter IV, we discuss our investigations of the reliability of Arbitron Radio audience estimates completed to date and the implications of the results of these investigations for the users of Arbitron Radio audience estimates.

Finally, in Chapter V, we present further details of the methodological and statistical procedures utilized in the study.

A Comprehensive Discussion of the Study

B. The Concepts Involved and Their Meaning

1. Sampling Error and Effective Sample Bases (ESB's)

When less than the full population is surveyed regarding television viewing or any other type of behavior, the results of the survey are surrounded by some degree of error due to the sampling process.

For obvious economic and practical reasons, we can never survey the total population. So we have to survey *samples* of the population. By the use of established statistical procedures we can determine the degree of sampling error present and thus the reliability of the results we obtain from samples.

Determining the size of the sampling error is a simple proposition if:

(1) Perfect simple random sampling (i.e., where each household in the population has an equal chance of being selected by the survey) is used; and

(2) Weighting households returning a usable viewing record to proportionately represent all household segments in the population is not necessary.

We would multiply the rating (p) as a percent for a station by its complement (q), which equals 100% − p, and divide the result by the total number of households (n) returning a usable (in-tab) viewing record. Then we would take the square root of the result.

The formula is as follows:

$$\textit{Standard Error of } (p) = \sqrt{\frac{pq}{n}}$$

The result of this calculation is referred to as the *Standard Error,* a statistical measure of sampling error.

For example, if we had a household audience rating of 5 (meaning that five percent of all television households in the survey area were viewing a station) and a resulting sample size of 400, the Standard Error of the rating (the margin of sampling error around the rating) would be calculated as follows:

$p = 5$

$q = 100\% - 5 = 95$

$n = 400$

$$\textit{Standard Error of } (p) = \sqrt{\frac{5 \times 95}{400}} = \sqrt{\frac{475}{400}} = \sqrt{1.188}$$

$$= 1.09$$

$$= \underline{1.1}$$

This indicates that with simple random sampling and no post-survey weighting, a rating of 5 calculated from the viewing records of 400 households is subject to a range of error of plus or minus 1.1 rating points, or a rating point range from 3.9 to 6.1. With this result, we can be about 68% sure that if we were able to measure the total population (using the same procedure), the rating for the total population would fall within this range.

If we wanted to be 95.5% sure that this would occur (the level most often used in broadcast audience research), we would multiply the Standard Error above (1.1) by *two,* which equals 2.2. The range of sampling error at this level of confidence would thus be plus or minus 2.2 rating points, or a range from 2.8 to 7.2.

The more confident we want to be of the reliability of the estimate, the wider the sampling error range.

Because it is not feasible to use simple random sampling in syndicated television audience surveys, and because we must weight in-tab samples to compensate for disproportionate diary returns from individual household groups, we cannot use the actual sample size (n) to calculate Standard Errors for resulting audience estimates.

Rather, we substitute a number referred to as the *Effective Sample Base (ESB)*, or effective sample size, for (n). The Effective Sample Base is defined as "the size of a simple random sample which would give the same standard error of an audience measurement as did the actual sampling plan upon which the result is based".[2] It indicates the size the survey sample is performing as, if simple random sampling and no post-survey weighting had been used.

The Effective Sample Base is the foundation factor for determining the magnitude of sampling error involved in syndicated broadcast audience estimates. The size of the effective sample influences the size of the Standard Error around the estimate and in turn influences the reliability of the estimate. For a given audience estimate, the larger the ESB, the smaller the Standard Error, and the more reliable the estimate.

The formula for computing the Standard Error for broadcast audience ratings is thus:

$$\textit{Standard Error of } (\mathrm{p}) = \sqrt{\frac{\mathrm{p\,q}}{\mathrm{ESB}}}$$

[2] *Standard Errors and Effective Sample Sizes as Reported for Broadcast Audience Measurement Surveys*, a publication by the Broadcast Rating Council, Inc., 1970, p. 21.

The Effective Sample Base is substituted for the in-tab sample size to get (n).

In the present Arbitron Television Market Report processing system, one Effective Sample Base for the total in-tab household sample is calculated. This ESB is calculated from a general theoretical formula which measures the degree of disproportionality in the distribution of household segments between the in-tab sample and population parameters, whether influenced by disproportionate diary returns or the initial sample design.

Under present Arbitron Television procedures, the reported ESB is an *approximation* of the actual ESB, calculated, as we have said, from a general theoretical formula. The goal of the study which Arbitron has been conducting over the past three years was to determine more precisely the size of the *true* ESB's for households and the various sex/age groups reported upon in Arbitron Television Market Reports. It is only through more precise knowledge of ESB's that we can assess more accurately the reliability of audience estimates.

2. Statistical Efficiency

Determining more precisely the size of true ESB's for audience estimates is similar in *principle* to the present ESB estimation procedure used by Arbitron Television, outlined in the previous section. It involves adjusting the actual in-tab sample size (n) to reflect the performance of the sample in representing total households and sex/age groups. But rather than being calculated from a theoretical formula, the adjustment is made on the basis of an efficiency factor developed from actual empirical data.

It is this factor, known as the *Statistical Efficiency* (SE), which is the key determinant of the ESB. As such, it is this factor which has been investigated over the past three years to arrive at conclusions regarding the size of true ESB's and thus the reliability of Arbitron Television audience estimates.

Applying the Statistical Efficiency factor to calculate the ESB is a simple process. We merely multiply the sample size (n) by the efficiency factor (SE):

$$\mathrm{ESB} = \mathrm{n} \times \mathrm{SE}$$

To calculate a Standard Error, we use this ESB value in the formula presented in the previous section:

$$\textit{Standard Error of } (\mathrm{p}) = \sqrt{\frac{\mathrm{p\,q}}{\mathrm{ESB}}} \text{ or } \sqrt{\frac{\mathrm{p\,q}}{\mathrm{n} \times \mathrm{SE}}}$$

Determining Statistical Efficiency values is a much more involved process, however. In the next section, we discuss the development of these values.

C. The Procedures Used in the Collection and Analysis of the Data

1. Developing Statistical Efficiency Values from Empirical Data

Statistical Efficiencies (SE's) were developed through the *replicated subsamples,* or *replication,* procedure. Statistical Efficiencies were developed only for ratings, but the findings for ratings are applicable to audience projections. The complete replication process was carried out as follows.

1. A national sample of usable designated households was drawn from the February/March 1972 Arbitron Television nationwide survey and divided randomly into five mutually exclusive subsamples, or replicates.[3] (See Appendix A for a detailed description of the procedure used to draw the national sample.)

2. Audience estimates for each of the five subsamples were processed using in-tab diaries returned from the usable designated homes.

3. Audience estimates for each subsample were developed using processing procedures identical to those used for published Arbitron Television Market Reports.

4. Audience estimates for each subsample were processed independently from the published February/March 1972 Television Market Report processing. Household weights were recalculated in the processing of audience estimates

[3] Within limits, the more subsamples used, the better the estimate of reliability. From a statistical and a practical standpoint, we determined that five subsamples was the most reasonable to use.

for each subsample; household weights from the published Market Report were not carried over to the audience estimates for each subsample.

5. Audience estimates for each subsample were calculated for five station categories:

   a. ABC Affiliates
   b. CBS Affiliates
   c. NBC Affiliates
   d. Independent Stations
   e. Educational Stations

6. For each rating from the subsamples, an arithmetic average rating ($\bar{p}$) across all five replicates was calculated.

7. For each of these average ratings ($\bar{p}$), a *Benchmark Variance* was calculated.

The *Benchmark Variance* is simply the Standard Error squared, or:

$$\textit{Benchmark Variance of } (\bar{p}) = \frac{\bar{p}q}{n}$$

Where,

$\bar{p}$ = average rating as a percent across all replicates
$q = 100\% - \bar{p}$
n = actual sample size across all replicates

This statistic indicates the degree of sampling error which would be present if the rating had been derived from a simple random sample and if no post-survey weighting had been used. It reflects, in other words, the hypothetical or theoretical situation.

8. For each of the average ratings ($\bar{p}$), an *Actual Variance* was also computed.

The *Actual Variance* is determined by applying the average rating ($\overline{p}$) and the ratings for each of the replicates in the following formula:

$$\textit{Actual Variance of } (p) = \frac{\sum_{i=1}^{m} (p_i - \overline{p})^2}{m\ (m-1)}$$

Where,

$\overline{p}$ = average rating as a percent across all replicates

p_i = individual rating as a percent from each replicate

$_i$ = the individual replicate, 1 to 5

m = number of replicates in total, which equals 5

This statistic reflects the *true* amount of variability among the random parts (replicates) of the total sample, rather than the hypothetical situation.

Since this measure of sampling error is computed using actual empirical data, any assumptions about the effects of sample clustering or stratification and post-survey weighting are avoided. All of the factors which influence sampling error, over and above the actual sample size, are automatically taken into account.

In addition, any processing errors in recording, editing, coding, and tabulating responses are automatically taken into account.

9. For each of the average ratings ($\overline{p}$), a *Statistical Efficiency* factor (SE) was calculated.

The *Statistical Efficiency* is the ratio of the *Benchmark Variance* or sampling error of the average rating ($\overline{p}$) to the *Actual Variance* or sampling error of the average rating ($\overline{p}$):

$$\textit{Statistical Efficiency } (\overline{p}) = \frac{\text{Benchmark Variance of } (\overline{p})}{\text{Actual Variance of } (\overline{p})}$$

10. For each of the average ratings ($\overline{p}$) for households and all reported sex/age groups within all reported time periods and dayparts, an *Effective Sample Base* (ESB) was calculated.

As discussed earlier, the ESB is calculated by multiplying the actual in-tab sample size (n) by the Statistical Efficiency factor (SE):

$$\textit{Effective Sample Base} = \begin{array}{l}\text{Actual In-Tab Sample Size}\\ \times\ \text{Statistical Efficiency } (\overline{p})\end{array}$$

or,

$$\text{ESB} = \text{n} \times \text{SE}$$

An Example

Before going further in our discussion, let us now pause to consider an example of the above steps, since we have now determined how the Statistical Efficiency factor is calculated and applied.

From the audience estimates for each of the replicates, we find the following household ratings for one network affiliate during the 7:30-11:00 PM, Sunday-Saturday, daypart:

Replicate (Subsample)					
1 (p_1)	2 (p_2)	3 (p_3)	4 (p_4)	5 (p_5)	Arithmetic Average ($\overline{p}$)
17.2	15.5	16.9	16.2	16.2	16.42

The total in-tab sample size (n) for households is 6,359.

First of all, to compute the Benchmark Variance, we multiply the average rating ($\overline{p}$) [16.42] by (q) [100% $-\ \overline{p}$ = 100.00% − 16.42 = 83.58] and divide by (n) [6,359]:

$$\textit{Benchmark Variance of } (\bar{p}) = \frac{\bar{p}q}{n}$$

$$= \frac{16.42 \times 83.58}{6,359}$$

$$= \underline{0.216}$$

Secondly, to compute the *Actual Variance*, we apply the average rating ($\bar{p}$) and the ratings for each replicate (p_i) to the formula presented below:

Actual Variance of ($\bar{p}$)

$$= \frac{\sum_{i=1}^{m} (p_i - \bar{p})^2}{m \ (m-1)}$$

(m = number of replicates = 5)

$$= \frac{(17.2\text{-}16.42)^2 + (15.5\text{-}16.42)^2 + (16.9\text{-}16.42)^2 + (16.2\text{-}16.42)^2 + (16.2\text{-}16.42)^2}{5\,(4)}$$

$$= \frac{0.6084 + 0.8464 + 0.2304 + 0.0484 + 0.0484}{20}$$

$$= \frac{1.7820}{20}$$

$$= \underline{0.089}$$

Thirdly, to compute the *Statistical Efficiency*, we divide the *Benchmark Variance* by the *Actual Variance:*

$$\textit{Statistical Efficiency}\ (\overline{\mathrm{p}}) = \frac{\text{Benchmark Variance of } (\overline{\mathrm{p}})}{\text{Actual Variance of } (\overline{\mathrm{p}})}$$

$$= \frac{0.216}{0.089}$$

$$= \underline{2.4}$$

Finally, to compute the ESB, we multiply the actual in-tab sample size (n) by the Statistical Efficiency factor (SE):

$$\mathrm{ESB} = \mathrm{n} \times \mathrm{SE}$$

$$= 6{,}359 \times 2.4$$

$$= \underline{15{,}262}$$

What this *Statistical Efficiency* factor tells us is that, *based on actual empirical data*, the sample of 6,359 households for this particular rating is performing as if it were 2.4 times larger than it really is. Or in terms of *ESB*, the sample of 6,359 households is performing as if it were an unweighted simple random sample of 15,262 households.

2. Analyzing Statistical Efficiencies

The calculation of Statistical Efficiencies outlined in the previous section was carried out for the thousands of specific Arbitron Television ratings resulting from the replication process.

The problem now was to apply these results to develop a system or model to determine a Statistical Efficiency factor for individual published Arbitron Television audience estimates without having to repeat the replication process. Without such a model, there would be no way to implement the results of the replication process to compute actual ESB's of Arbitron Television audience estimates. It would simply not be feasible to repeat the replication process to determine empirically individual Statistical Efficiencies.

The first step in the model building was to analyze the results of the replication procedure to determine what variable or variables influence the size of the Statistical Efficiency factor. Among the variables studied were: Size of estimate, day of week, time of day, length of daypart, station, station affiliation, and demographic group.

From extensive analyses, it was found that there are two key *interrelated* determinants of the size of the Statistical Efficiency factor:

(1) **The number of quarter-hours averaged to develop the audience estimate.**

As the number of quarter-hours averaged to develop an audience estimate *increases*, the Statistical Efficienc *increases*.

An explanation of this finding is as follows.

An audience estimate for one quarter-hour is based upon a single observation or sample of the total respondent sample. An estimate for more than one quarter-hour is based upon more than one observation of the total respondent sample. The more observations made before the data are combined, the more stable the average will be—and in turn, the higher the Statistical Efficiency will be.

Of course, as the Statistical Efficiency increases so does the Effective Sample Base. And for a given audience estimate, the larger the ESB, the smaller the sampling error around that estimate and thus the more reliable it is.

As observations of the same total respondent sample are repeated, the gain in Statistical Efficiency occurs at a declining rate. This is because continued observations of the same sample contribute less new information each time a new observation is made. However, as more observations are made and combined, some new information is gained which contributes to more reliable estimates of viewing behavior.

(2) **The population group upon which the audience estimate is based.**

Smaller, more tightly-defined demographic groups tend to have higher Statistical Efficiencies than larger, less tightly-defined demographic groups.

This is because we are more efficient in a statistical sense in measuring the viewing behavior of smaller, more tightly-defined demographic groups.

The tighter the demographic definition of the population group, the less likely we are to find two or more people in this group who live in the same household. Thus, there is less "clustering" effect in repeated observations of viewing in the same household.

In addition, smaller and more tightly-defined demographic groups are somewhat more likely to view television during the same time period. Thus, there is more efficiency in measuring these groups' viewing behavior through survey sampling.

3. The Statistical Efficiency Model

With the knowledge of the key determinants of the Statistical Efficiency value, the next step was to apply this knowledge and the specific results of the replication study to develop a mathematical model (or models), applicable to all markets, which could predict Statistical Efficiency values precisely enough for implementation with Arbitron Television audience estimates.

The requirement for the modelling procedure was that it enable the data user to determine accurately the Statistical Efficiency value to be used in calculating the sampling error for an audience estimate given the known information:

(1) The number of quarter-hours averaged to compute the audience estimate; and

(2) The population group upon which the audience estimate is based.

After a considerable amount of investigation and analysis, a general model of Statistical Efficiencies was developed by

MarketMath. The general model is built upon two separate models which were applied in two distinct steps.

The first step involved relating the number of quarter-hours averaged (to compute an audience estimate) to Statistical Efficiency values *within* individual demographic groups. The resulting model is a rational statistical model, derived on logical grounds, which when tested against observed data results in an extremely close fit (i.e., predicts Statistical Efficiency values extremely well).

This model was applied separately to each demographic group included in the replication analysis, since the relation of Statistical Efficiency to the number of quarter-hours averaged varies by population group, as noted in the previous section. For a technically-oriented discussion of this model, see Appendices B and C.

The second step in the general model of Statistical Efficiencies involved smoothing Statistical Efficiencies *across* demographic groups to adjust for slight differences between modelled and observed values and to estimate Statistical Efficiencies for demographic groups not covered in the replication analysis. This was done using an empirical regression model, which is discussed in Appendix B.

Resulting from the modelling procedures was a table of Statistical Efficiency values covering all time periods and population groups reported upon in current Arbitron Television Market Reports. This table is presented in Chapter III in conjunction with the guide to calculating more accurate Standards Errors of Arbitron Television audience estimates.

Chapter III

Implementing the Results

Implementing the Results

A. A Guide to Calculating More Accurate Standard Errors of Arbitron Television Audience Estimates

Calculating More Accurate Standard Errors of Arbitron Television Audience Estimates

In this section, we discuss the most important aspect of this study—the implementation of the results of the study in calculating more accurately the Standard Error of any audience estimate published in current Arbitron Television Market Reports.

Note that this procedure applies only to ratings, rating sums, and HUT's. To calculate the Standard Error of projections or shares, these estimates must first be converted to a percentage (or rating) basis.

This procedure involves seven steps:

1. Determine the rating (p) for the station, population group, and time period or daypart in question from the Market Report.

2. Subtract the rating (p) from 100% to determine (q), the complement of the rating.

3. Determine the survey in-tab sample size (n) for the population group upon which the rating is based.

   Household in-tab sample sizes for the market's Metro Area, Area of Dominant Influence (ADI), and Total Survey Area (TSA) are reported on page seven of the Television Market Report. These numbers represent (n) for Standard Error calculations for HUT's and household ratings.

In-tab sample sizes for sex/age groups in the market's ADI and TSA are also reported on page seven of the Television Market Report. These numbers represent the (n) for Standard Error calculations for any demographic ratings.

4. Determine the number of quarter-hours averaged to calculate the rating (p).

This is accomplished by multiplying the number of quarter-hours in the time period or daypart each day by the number of days in the daypart.

5. Determine the Statistical Efficiency factor (SE) from the table presented on page 54.

Find the population group in question in the lefthand column of the table. Then follow the row of numbers to the right of this column until you reach the column for the number of quarter-hours in the time period or daypart in question.

For your reference, we have provided on page 55, a chart which shows the number of quarter-hours averaged in the time periods and daypart estimates published in current Arbitron Television Market Reports.

6. Enter the numbers determined above in the formula for the Standard Error:

$$\textit{Standard Error of } (\mathrm{p}) = 2 \times \sqrt{\frac{\mathrm{pq}}{\mathrm{ESB}}} = 2 \times \sqrt{\frac{\mathrm{pq}}{\mathrm{n} \times \mathrm{SE}}}$$

NOTE: The Standard Error at the 95.5% level of confidence (i.e., the Standard Error is multiplied by two) is used here since this is the most accepted and used level of confidence.

7. Determine the confidence interval for the rating (p).

This is accomplished by subtracting the resulting Standard Error from (p) and adding the same value to (p).

This indicates the range within which we can be 95.5% certain that the specific Television rating in question would fall if we measured the total population from which the sample was drawn.

An Example

As an example, consider an ADI station rating of 10 for the Women 18-49 sex/age category during the 4-6:30 PM, Monday-Friday, daypart. The Standard Error calculation is as follows:

(1) Market Report rating (p) = $\underline{10}$

(2) (q) = $\underline{90}$

$q = 100\% - p = 100\% - 10 = 90$

(3) In-Tab Sample Size (n) = $\underline{300}$

(4) Number of quarter-hours averaged = $\underline{50}$

4-6:30 PM, Monday-Friday, covers two and a half hours or ten quarter-hours per day for five days.

$10 \times 5 = 50$

(5) Statistical Efficiency (SE) = 2.4

From the table, the SE value for the Women 18-49 sex/age category and 50 averaged quarter-hours is 2.4.

(6) Standard Error of (p)

$$= 2 \times \sqrt{\frac{pq}{n \times SE}}$$

$$= 2 \times \sqrt{\frac{10 \times 90}{300 \times 2.4}}$$

$$= 2 \times \sqrt{\frac{900}{720}}$$

$$= 2 \times \sqrt{1.250}$$

$$= 2 \times 1.118$$

$$= 2.24$$

$$= \underline{2.2}$$

(7) The 95.5% confidence interval for the rating is 10.0 ± 2.2 or 7.8 to 12.2.

This indicates that we can be 95.5% certain that the *true* percentage of the population of Women 18-49 in an ADI viewing this station during this daypart falls within this range.

Had we calculated the Standard Error of this rating using the traditional "nomograph" procedure, the Standard Error would equal 3.5. The 95.5% confidence interval would thus be 10.0 ± 3.5, or 6.5 to 13.5. Using the more accurate procedure, the Standard Error and confidence interval are reduced by 37%!

Statistical efficiencies for population groups by number of quarter-hours in a time period or daypart

	1	2	5	10	18	30	32	50	60	70	84	90	98	420
Households	.9	.9	1.1	1.4	1.8	2.3	2.4	3.2	3.6	4.0	4.4	4.6	4.9	10.7
Persons 2+	.4	.4	.5	.6	.8	1.1	1.2	1.7	1.9	2.2	2.5	2.7	2.9	11.0
Persons 12-34 ...	.5	.6	.9	1.4	2.0	2.8	2.9	3.9	4.3	4.6	5.1	5.3	5.5	8.5
Adults 18+	.5	.6	.8	1.1	1.6	2.2	2.3	3.1	3.6	3.9	4.5	3.9	4.9	9.9
Total Women	.6	.8	1.1	1.5	1.9	2.2	2.2	2.5	2.6	2.6	2.7	2.7	2.7	3.0
Women 18-49 ...	.6	.8	1.0	1.5	1.8	2.1	2.2	2.4	2.5	2.6	2.6	2.7	2.7	3.0
18-34 ...	.6	.8	1.1	1.4	1.8	2.2	2.3	2.5	2.6	2.7	2.8	2.8	2.9	3.2
25-64 ...	.7	.8	1.1	1.5	1.9	2.2	2.2	2.5	2.6	2.6	2.7	2.7	2.8	3.1
25-49 ...	.7	.8	1.1	1.4	1.8	2.1	2.2	2.4	2.5	2.6	2.7	2.7	2.7	3.0
50+ ...	.8	.9	1.2	1.6	2.1	2.6	2.7	3.2	3.3	3.5	3.6	3.7	3.7	4.4
Total Housewives .	.7	.8	1.1	1.5	1.9	2.2	2.2	2.5	2.5	2.6	2.7	2.7	2.7	3.0
Total Men	.7	.8	1.1	1.6	2.3	3.2	3.3	4.4	4.9	5.3	5.9	6.1	6.4	10.7
Men 18-49	.7	.8	1.1	1.6	2.3	3.2	3.4	4.7	5.3	5.9	6.7	7.0	7.4	15.0
18-34	.7	.8	1.0	1.5	2.1	3.0	3.1	4.2	4.8	5.3	5.9	6.2	6.5	12.8
25-64	.7	.8	1.1	1.5	2.2	3.1	3.3	4.5	5.2	5.8	6.6	6.9	7.3	15.7
25-49	.7	.8	1.1	1.5	2.1	3.0	3.2	4.4	5.1	5.7	6.6	6.9	7.3	18.4
Total Teens	.5	.6	.8	1.1	1.5	2.0	2.1	2.7	3.0	3.2	3.5	3.6	3.7	5.4
Girl Teens	.5	.7	1.0	1.4	1.8	2.2	2.3	2.6	2.7	2.8	2.8	2.9	2.9	3.3
Total Children ...	.3	.4	.6	.8	1.1	1.4	1.4	1.7	1.8	1.9	2.0	2.0	2.1	2.5
Children 6-11	.4	.5	.7	.9	1.3	1.7	1.8	2.2	2.4	2.6	2.8	2.8	2.9	4.0

Number of quarter-hours in time period and daypart estimates published in Arbitron Television Market Reports

Quarter-Hours	Time Period or Daypart
1	All single day quarter-hour time periods ex., 11-11:15 pm, Tuesday
2	All single day half-hour time periods ex., 9-9:30 pm, Thursday
5	All Monday-Friday quarter-hour time periods ex., 5-5:15 pm, Monday-Friday
10	All Monday-Friday half-hour time periods ex., Noon-12:30 pm, Monday-Friday 6-6:30 pm, Monday-Friday daypart 7-7:30 pm, Monday-Friday daypart 6:30-7 pm, Monday-Friday daypart 7:30-8 pm, Monday-Friday daypart 9:30-10 pm, Monday-Friday daypart 10:30-11 pm, Monday-Friday daypart 10-10:30 pm, Monday-Friday daypart 11-11:30 pm, Monday-Friday daypart
18	8:30 am-1 pm, Saturday daypart
30	3:30-5 pm, Monday-Friday daypart 4:30-6 pm, Monday-Friday daypart 5-6:30 pm, Monday-Friday daypart 6-7:30 pm, Monday-Friday daypart 10:30 pm-Midnight, Monday-Friday daypart 11:30 pm-1 am, Monday-Friday daypart
32	1-5 pm, Saturday + Sunday daypart
50	4-6:30 pm, Monday-Friday daypart 5-7:30 pm, Monday-Friday daypart
60	9 am-Noon, Monday-Friday daypart
70	6:30-10 pm, Monday-Friday daypart 7:30-11 pm, Monday-Friday daypart
84	7-10 pm, Sunday-Saturday daypart 8-11 pm, Sunday-Saturday daypart
90	11 am-3:30 pm, Monday-Friday daypart Noon-4:30 pm, Monday-Friday daypart
98	6:30-10 pm, Sunday-Saturday daypart 7:30-11 pm, Sunday-Saturday daypart
420	9 am-Midnight, Sunday-Saturday daypart

B. A Guide to Determining if Audience Estimate Differences are Statistically Significant

Determining if Audience Estimate Differences are Statistically Significant

In the previous section, the procedure for calculating more accurately the Standard Error of an individual audience estimate was explained. Because one of the goals of this study was to investigate the fluctuation in audience estimates from survey-to-survey, we will now explain how to apply the numbers used to calculate the Standard Error of individual audience estimates to determine if the difference between two estimates for the same time period or daypart and population group, but separate surveys, is statistically significant.

There are two steps involved in determining if the difference between two audience estimates (expressed in terms of ratings) is statistically significant:

1. Determine the Standard Error of the difference between the ratings.

The formula for accomplishing this, adapted for broadcast audience research usage, is as follows:

$$\textit{Standard Error of Difference in Ratings } (p_1, p_2) = \sqrt{\frac{p_1q_1}{ESB_1} + \frac{p_2q_2}{ESB_2}}$$

Where,

p_1 = rating as a percent from one survey

p_2 = rating as a percent for same time period or daypart and population group from another survey

q_1 = $100\% - p_1$

q_2 = $100\% - p_2$

ESB_1 = Effective Sample Base for one survey; ESB equals in-tab sample size (n) for this survey multiplied by the Statistical Efficiency value (SE)

ESB_2 = Effective Sample Base for the second survey; ESB equals in-tab sample size (n) for this survey multiplied by the Statistical Efficiency value (SE)

To calculate the Standard Error of the difference between two ratings, we work with the same numbers used to calculate the true Standard Error of the individual ratings as explained in the previous section. We simply take the value of the true Standard Error of the individual rating *before its square root is computed*, add the values for the two individual ratings, and then take the square root of the sum.

2. Determine the criterion point at which the difference between the two ratings (p_1 and p_2) becomes statistically significant.

To accomplish this we multiply the Standard Error of the difference between the two ratings (p_1 and p_2) by *two:*

Criterion Point = Standard Error of Difference in Ratings (p_1, p_2) × 2

This gives us a criterion value for deciding with 95.5% confidence whether a rating point difference between two sample surveys denotes a real change in the population being sampled. The number *two* indicates the number of Standard Errors necessary to be sure that a rating difference between two sample surveys denotes a real change. It is comparable in principle to our multiplying the Standard Error of the individual rating by *two* to compute a 95.5% confidence interval.

If the difference between two ratings is *larger than* or *equal to* this criterion value, we can be 95.5% sure that the larger of the two sample ratings would remain larger if we increased the sample size indefinitely while maintaining all other survey methods.

An Example

As an example of how this procedure is carried out, consider again the example from the previous section:

An ADI rating for Women 18-49, 4-6:30 PM, Monday-Friday daypart, November survey:

Market Report Rating $(p_1) = \underline{10}$

$$\begin{aligned} q_1 &= \underline{90} \\ ESB_1 &= n \times SE \\ &= 300 \times 2.4 \\ &= \underline{720} \end{aligned}$$

Now consider the rating for the same ADI, the same sex/age group, and the same daypart, but for the following February/March survey:

Market Report Rating $(p_2) = \underline{7}$

$$q_2 = \underline{93}$$

$$\begin{aligned} ESB_2 &= n \times SE \\ &= 320 \times 2.4 \\ &= \underline{768} \end{aligned}$$

Since the November rating was determined before the February/March rating, we have labelled it the first rating (p_1). The February/March rating is labelled the second rating (p_2). Note that the difference between the two ratings is 3 points $(p_1 = 10$ vs. $p_2 = 7)$. Our goal here then is to determine if the rating (p_2) is in truth lower than the rating (p_1).

We will first calculate the Standard Error of the difference between the two ratings $(p_1$ and $p_2)$. Applying the numbers above to the formula, we have:

Standard Error of Difference in Ratings (p_1, p_2) $= \sqrt{\frac{p_1q_1}{ESB_1} + \frac{p_2q_2}{ESB_2}}$

$$= \sqrt{\frac{10 \times 90}{720} + \frac{7 \times 93}{768}}$$

$$= \sqrt{\frac{900}{720} + \frac{651}{768}}$$

$$= \sqrt{1.250 + 0.848} = \sqrt{2.098}$$

$$= \underline{1.45}$$

Now to determine the criterion point, we multiply 1.45 by two:

$$\begin{aligned} \textit{Criterion Point} &= 1.45 \times 2 \\ &= \underline{2.9} \end{aligned}$$

This tells us that to be 95.5% certain that the rating (p_2) is in truth lower than the rating (p_1), the rating (p_2) would have to be at least 2.9 points lower than the rating (p_1). Since the actual difference is 3 points, we can conclude that the rating for the February/March survey is in truth lower than that for the November survey.

Chapter IV

Progress in the Investigation of the Reliability of Arbitron Radio Audience Estimates

Progress in the Investigation of the Reliability of Arbitron Radio Audience Estimates

A. Summary—Apparent Implications

Summary—Apparent Implications

Although the primary focus of this report is on the reliability of Arbitron Television audience estimates, we should not fail to mention the research completed to date on the reliability of Arbitron Radio audience estimates.

Thus far, we have completed an analysis of the reliability of Arbitron Radio audience estimates for two individual markets. The study has not progressed as yet to a national sample analysis, but such an analysis is planned.

In brief, what the study of Arbitron Radio audience estimates has shown is that all estimates, other than cume estimates, are more reliable or precise than indicated by current approximation procedures. Resulting Statistical Efficiency values, except those for cumes, are greater than one, showing that the true Effective Sample Base for most Arbitron Radio audience estimates is greater than the in-tab sample size. This means that Arbitron Radio survey samples are performing as if they were larger than the simple total of respondents in the sample would imply.

This conclusion, as you will recognize, is basically the same as the one reached for Arbitron Television estimates. However, the apparent implications of this conclusion are somewhat more dramatic for Radio estimates.

A considerable number of published Arbitron Radio audience estimates are based on relatively small population groups, and the sample sizes for these estimates are necessarily small. However, because most of the published estimates for these groups are based upon a large number of averaged quarter-hours, their Statistical Efficiencies are high, relatively speaking, and the estimates are much more reliable than currently used approximation formulas would lead us to believe.

Thus, Radio audience estimates for even the smallest demographic groups are more reliable and sensitive descriptors of audience size than some would think. As a result, a station's performance can now be evaluated across many demographics with more certainty that the performance is being measured precisely.

Progress in the Investigation of the Reliability of Arbitron Radio Audience Estimates

B. The Radio Estimates Reliability Study

The Radio Estimates Reliability Study

Our goal in the study of the reliability of Arbitron Radio audience estimates was to determine how precisely or reliably sample measurement of Radio listening represents listening by the total population from which the sample is drawn.

As for Arbitron Television audience estimates, the precision of Arbitron Radio audience estimates (expressed in terms of Standard Errors) is presently calculated using a formula which assumes that the estimates are derived from an unweighted simple random sample. Since this is not the true situation involved in the survey, the resulting Standard Error is at best a rough *approximation* of the precision of the estimates.

To calculate more accurate Standard Errors of Arbitron Radio audience estimates, we again used the replication procedure.

Following a recommendation by the National Association of Broadcasters (NAB), with which we worked cooperatively in designing the Radio study, the replication procedure was carried out using the total in-tab sample from the October 1971 Arbitron Radio survey in the Indianapolis and Philadelphia Metro areas.

The applicable concepts, methodology, and supporting mathematics involved in the replication process for these individual Radio markets are essentially the same as those discussed earlier for the national Television replication project, so we need not repeat them here.

The thousands of Statistical Efficiency values resulting from the replication procedure were analyzed to determine how to maximize their utility for Arbitron Radio audience data users. Through these analyses, we discovered that the two key interacting variables which influence the Statistical Efficiency of Television audience estimates (as presented in the next paragraph) are also the key determinants of the Statistical Efficiency of Radio audience estimates.

Likewise, we discovered that the general model of Statistical Efficiencies developed from the national Television data is applicable to Radio data. The general model is capable of predicting the Statistical Efficiency of an Arbitron Radio audience estimate given the known information:

(1) the number of quarter-hours averaged to compute the audience estimate; and

(2) the population group upon which the audience estimate is based.

The model was applied to calculate a table of Statistical Efficiency values which can be used to compute more accurately the Standard Error, or reliability, of any audience estimate published in Arbitron Radio Market Reports, knowing only the two variables discussed in the previous paragraph. The table of Radio Statistical Efficiencies is presented in the next section, along with a guide for its use in the calculation of Standard Errors. As presented, the table does not contain Statistical Efficiency values for all dayparts. When our investigations are complete, we will provide Statistical Efficiencies applicable to all dayparts.

Although the table is based on data from only two individual Radio markets, we feel it provides valuable information, because it is indicative of what we believe can be expected when further analyses are made based on a national sample. Through our investigations of a national sample, we plan to test the data in the table of Radio Statistical Efficiencies to provide a more definitive understanding of the reliability of these estimates.

Progress in the Investigation of the Reliability of Arbitron Radio Audience Estimates

C. A Guide to Calculating More Accurate Standard Errors of Arbitron Radio Audience Estimates

Calculating More Accurate Standard Errors of Arbitron Radio Audience Estimates

In this section, we describe how the results of our study of the reliability of Arbitron Radio audience estimates can be implemented to calculate more accurately the Standard Error of any audience estimate published in current Arbitron Radio Market Reports. The procedure is essentially the same as the one described for Television estimates in Chapter III; however, a few of the mechanics differ.

Note that this procedure applies only to ratings and rating sums. To calculate the Standard Error of projections or shares, these estimates must first be converted to a percentage (or rating) basis. This procedure is then applicable to both the Metro and Total Survey Area audience estimates.

This procedure involves seven steps:

(1) Determine the rating (p) for the station, population group, and daypart in question from the Market Report;

(2) Subtract the rating (p) from 100% to determine (q), the complement of the rating;

(3) Determine the survey in-tab sample size (n) for the population group upon which the rating (p) is based.

This is accomplished by multiplying the percentage for the population group under the column "Percent of Unweighted In-Tab Sample" by the total in-tab sample under the column "Total Tabulated Diaries", both of which are shown on page three of each Radio Market Report.

(4) Determine the number of quarter-hours averaged to calculate the rating (p).

This is accomplished by multiplying the number of quarter-hours in the daypart each day by the number of days in the daypart.

(5) Determine the Statistical Efficiency factor (SE) from the table presented at the end of this section.

Find the population group in question in the left-hand column of the table. Then follow the row of numbers to the right of this column until you reach the column for the number of quarter-hours in the daypart in question.

For your reference, we have provided a chart which shows the specific time periods and dayparts covered by the table of Radio Statistical Efficiencies.

(6) Enter the numbers determined above in the formula for the Standard Error:

$$\textit{Standard Error of } (\mathrm{p}) = 2 \times \sqrt{\frac{\mathrm{pq}}{\mathrm{ESB}}} = 2 \times \sqrt{\frac{\mathrm{pq}}{\mathrm{n} \times \mathrm{SE}}}$$

NOTE: The Standard Error at the 95.5% level of confidence (i.e., the Standard Error is multiplied by two) is used here since this is the most accepted and used level of confidence.

(7) Determine the confidence interval for the rating (p). This is accomplished by subtracting the resulting Standard Error from (p) and adding the same value to (p).

This indicates the plus-minus range within which we can be 95.5% certain that the specific Radio rating in question would fall if we measured the total population from which the sample was drawn.

An Example

As an example, consider the Metro station rating of 5.9 for the Men 35-49 sex/age category during the 6-10 AM, Monday-Friday, daypart in the Boston April/May 1973 Radio Market Report. The Standard Error calculation is as follows:

(1) Market Report rating (p) = $\underline{5.9}$

(2) (q) = $\underline{94.1}$

$q = 100.0\% - p = 100.0\% - 5.9 = 94.1$

(3) In-tab Sample Size (n) = $\underline{92}$

$n = .093 \times 987 = 91.8 = 92$

(4) Number of quarter-hours averaged = $\underline{80}$

6-10 AM, Monday-Friday, covers four hours or 16 quarter-hours per day for five days.

$16 \times 5 = 80$

(5) Statistical Efficiency (SE) = $\underline{3.1}$

From the table, the SE value for the Men 35-49 sex/age category and 80 averaged quarter-hours is 3.1.

(6) Standard Error of (p)

$$\begin{aligned}
&= 2 \times \sqrt{\frac{pq}{n \times SE}} \\
&= 2 \times \sqrt{\frac{5.9 \times 94.1}{92 \times 3.1}} \\
&= 2 \times \sqrt{\frac{555.19}{285.20}} \\
&= 2 \times \sqrt{1.947} \\
&= 2 \times 1.395 \\
&= 2.79 \\
&= \underline{2.8}
\end{aligned}$$

(7) The 95.5% confidence interval for the rating 5.9 is ± 2.8, or 3.1 to 8.7.

This indicates that we can be 95.5% certain that the true rating for the population of Men 35-49 in the Boston Metro for this station and daypart falls within this range.

Had we calculated the Standard Error of this rating using the conventional "nomograph" procedure, the standard Error would equal 5.3. The 95.5% confidence interval would thus be 5.9 ± 5.3, or 0.6 to 11.2. Using the more accurate procedure, the Standard Error and confidence interval are reduced by 47%!

Radio statistical efficiencies for population groups by number of quarter-hours in a time period or daypart

		Cume					
		Ratings	20	80	100	160	504
Total Persons	12+	.5	1.2	1.9	2.0	2.1	2.4
Total Adults	18+	.6	1.3	2.0	2.1	2.3	2.6
Total Men	18+	.7	1.4	2.5	2.8	3.2	4.1
Total Women	18+	.7	1.5	2.5	2.7	3.0	3.4
Adults	25-64	.7	1.4	2.3	2.4	2.7	3.1
Men	25-64	.7	1.4	2.7	3.0	3.5	4.6
Women	25-64	.7	1.5	2.7	2.9	3.2	3.8
Adults	18-49	.6	1.5	2.6	2.7	3.1	3.6
Men	18-49	.7	1.5	3.0	3.4	4.1	5.5
Women	18-49	.7	1.6	2.9	3.2	3.6	4.3
Adults	35-64	.7	1.4	2.4	2.6	2.9	3.4
Men	35-64	.7	1.4	2.8	3.1	3.7	4.9
Women	35-64	.7	1.6	2.8	3.0	3.4	4.1
Adults	25-49	.7	1.5	2.7	2.9	3.3	3.9
Men	25-49	.7	1.5	3.0	3.4	4.2	5.7
Women	25-49	.7	1.6	3.0	3.2	3.7	4.5
Adults	50+	.8	1.5	2.6	2.7	3.1	3.8
Men	50+	.8	1.4	2.8	3.1	3.8	5.3
Women	50+	.8	1.6	2.8	3.0	3.5	4.2
Adults	35-49	.7	1.5	2.8	3.1	3.6	4.5
Men	35-49	.7	1.5	3.1	3.5	4.3	6.2
Women	35-49	.7	1.6	3.1	3.3	3.9	4.8
Adults	18-34	.6	1.6	3.2	3.5	4.1	5.2
Men	18-34	.7	1.6	3.7	4.2	5.3	7.8
Women	18-34	.7	1.7	3.4	3.7	4.4	5.5
Adults	50-64	.8	1.5	2.7	3.0	3.4	4.3
Men	50-64	.8	1.5	2.9	3.3	4.1	5.9
Women	50-64	.8	1.6	2.9	3.2	3.7	4.6
Teens	12-17	.6	2.0	4.4	4.9	5.9	7.9
Adults	25-34	.7	1.6	3.3	3.7	4.4	5.7
Men	25-34	.7	1.6	3.6	4.2	5.3	8.2
Women	25-34	.7	1.7	3.4	3.8	4.4	5.6
Adults	18-24	.6	1.8	3.9	4.3	5.2	7.0
Men	18-24	.6	1.7	4.3	4.9	6.3	9.9
Women	18-24	.6	1.9	3.9	4.3	5.1	6.6

Number of quarter-hours in selected time period and daypart estimates published in Arbitron Radio Market Reports

Quarter-Hours	Time Period or Daypart
Cume Ratings	All Cume Ratings regardless of time period, daypart, or daypart combination involved.
20	All Monday-Friday, one-hour time periods ex., 7-8 am, Monday-Friday 10 am-3 pm, Saturday daypart 10 am-3 pm, Sunday daypart 7 pm-Midnight, Saturday daypart 7 pm-Midnight, Sunday daypart
80	6-10 am, Monday-Friday daypart 3-7 pm, Monday-Friday daypart
100	10 am-3 pm, Monday-Friday daypart 7 pm-Midnight, Monday-Friday daypart
160	6-10 am + 3-7 pm, Monday-Friday daypart
504	6 am-Midnight, Monday-Sunday daypart

Chapter V

Appendices

Appendices

Appendix A. Procedure for Drawing the National Sample

1. To accomplish the processing of Television audience estimates for each replicate or subsample, each ADI (rank 1 to 149) was considered to be a single sample unit, and the ADI's rank 150 to 207 were grouped together as the 150th unit. In this way, only the smallest markets were grouped together.

The sample was then drawn to ensure a proportionate representation of all 207 ADI's within these 150 units.

2. Fifteen hundred sampling points were distributed among the 150 units proportionate to the number of television households in each. This step identified the specific counties, and the number of sampling points per county, from which the sample of households would be drawn.

3. It was estimated that approximately 12,000 usable designated households (those households receiving diaries) would be required to provide approximately 1,300 in-tab diaries in each replicate, or subsample. To achieve this total, eight usable designated households per sampling point were selected, using the existing sequence of sample households. This existing sequence of households was originally randomly selected and distributed over the county and the four week period in the February/March 1972 survey.

4. The entire specified household sample of 12,000 was then randomly distributed into five replicates or subsamples of 2,400 households.

5. The next step was to determine which of the 2,400 households in each subsample returned an in-tab diary during the February/March 1972 survey. The in-tab homes were then used to process complete Television audience estimates for the subsample.

The process involved here emulated the conduct of an actual survey, allowing each subsample to experience its own sample performance.

6. Each replicate ended up with approximately 1,300 in-tab diaries, as originally planned. The specific in-tab sample sizes were:

Replicate 1	1,304 households
Replicate 2	1,282 households
Replicate 3	1,310 households
Replicate 4	1,225 households
Replicate 5	1,238 households

Appendix

B. The General Model of Statistical Efficiencies

by Jerome D. Greene
MarketMath, Inc.

The general model of Statistical Efficiencies was developed to explain our replication results so that we might then estimate Statistical Efficiencies for all Arbitron Television ratings. For every demographic group and daypart included in the replication analysis, Statistical Efficiencies were estimated for each station-rating.

Each of these estimated Statistical Efficiencies is a variance ratio (the variance expected from simple random sampling divided by the variance estimated from replication), and each is subject to sampling error. We hypothesized that the differences in estimated Statistical Efficiency across stations within any one demographic group and daypart were due only to sampling error. That is, that the observed differences in estimated Statistical Efficiency across stations were entirely random. Therefore, a composite of Statistical Efficiencies across stations would be a better estimate of the true Statistical Efficiency for each station-rating than the one estimated for it by replication. We further hypothesized that as variance ratios, the Statistical Efficiencies estimated for each station-rating in any one demographic group and daypart would conform to the well-known theoretical distribution of variance ratios originally derived in logarithmic form by R. A. Fisher and later developed by G. W. Snedecor as the "F-distribution" (so named by Snedecor after Fisher).

Dr. Martin Frankel, consultant to the Broadcast Rating Council, has pointed out that if Statistical Efficiencies (based on 5 replications) are F-distributed, the proper average across stations is the harmonic mean (the reciprocal of the mean-reciprocal). Dr. Frankel used Design Effect (1/SE) in his analysis of the reliability of ratings, and averaged Design Effects across stations using the arithmetic mean ($\overline{DE} = 1/\overline{SE}$).

In our analysis, for each demographic group and daypart separately, we averaged Statistical Efficiencies across stations using the harmonic mean. Having done this, we were able to make extensive tests of the actual variance of Statistical Efficiencies (or Design Effects) across Stations against the variance we would expect from the F-distribution. Our results have confirmed the F-distribution hypothesis: The actual variance of Statistical Efficiencies across stations within each demographic group and daypart is generally no more than, and often less than, we would expect from the F-distribution. Then the differences in Statistical Efficiencies estimated for stations within any one demographic group and daypart are random. We are therefore entitled, and indeed ought to average Statistical Efficiencies across stations.

We observed that these average Statistical Efficiencies are lowest for cumulative ratings, and for average ratings increase as t, the number of averaged quarter-hours, increases. We have also observed that the Statistical Efficiencies have no tendency, apart from t, to vary by daypart.

Our next step was to develop a model to estimate Statistical Efficiencies separately for each demographic group as a function of the number of averaged quarter-hours (t). Here we refer you to Appendix C, a paper originally prepared in August 1972 and reprinted here. The paper is the rational basis for the "SET" model and should be read before continuing. It is important to note, however, that the paper deals with the relative gain in information (Statistical Efficiency) across increasing number of media units (in this case, quarter-hours), not the absolute gain, and does not therefore completely specify the model used to analyze our replication results.

Substituting C for $\frac{\sigma_p^2}{p \cdot q}$ in equation (5) of the paper, we get:

$$SE_t = \frac{t}{1 + (t - 1) C}$$

When $t = 1$, $SE_1 = 1$, as is the case in simple random sampling. To allow for stratification, clustering, and weighting effects, we add parameter D:

$$SE_t = \frac{Dt}{1 + (t - 1) C}$$

Now when $t = 1$, $SE_1 = D$, the "going-in" Statistical Efficiency. This in turn assumes that the Statistical Efficiency increases as t increases and that the increase is proportional to the "going-in" Statistical Efficiency.

But suppose, more realistically and conservatively, that the Statistical Efficiency increases with t only in proportion to a part of the going-in Statistical Efficiency. Then we should split parameter D into two components, A and B. This gives the final model:

$$SE_t = A + \frac{Bt}{1 + (t - 1)C}$$

Empirically, the use of parameter A greatly increases the goodness-of-fit.

The model was fit to the replication data for each demographic group separately to minimize:

$$\sum \frac{(Y - Y_c)^2}{Y}$$

where:

$Y = SE$ observed
$Y_c = SE$ by model

The "SET" model was used only for smoothing and interpolating; that is, the model was fit to and used to estimate Statistical Efficiencies for only those demographic groups and the range of t values included in the replication analysis.

Having fit the "SET" model to each demographic group separately, we faced two new related problems. First, for the same reasons we chose to average Statistical Efficiencies across stations and then further to smooth them across t values by the "SET" model, we wanted now to smooth them across demographic groups. Our second related problem was to estimate Statistical Efficiencies for those demographic groups not explicitly included in the replication analysis. Although all the demographic groups for which Arbitron then reported ratings were included in the replication analysis, at the request of clients, Arbitron has since begun to publish ratings for new demographic groups (different combinations of the "pieces" included in the replication analysis).

Our final step, then, was a Regression-smoothing model which takes into account t, the size, sex, and age composition of each demographic segment. This final model solves our last two problems. First, like the "SET" model which estimates Statistical Efficiencies as functions of t, it further smooths our results and reduces their chance variability. Second and more important, it allows us to estimate Statistical Efficiencies for demographic groups and dayparts which were not included in the replication analysis but might appear in Arbitron published reports.

In summary then, we have learned that the differences in estimated Statistical Efficiencies across stations are random, that we do not need separate estimates for each station. We know also that Statistical Efficiencies for average ratings are functions of the number of averaged quarter-hours and demographic variables. Accordingly, we can estimate with useful accuracy the Statistical Efficiency of any Arbitron daypart cume or average rating.

Appendix

C. A Note About the Information Gain from Interviewing a Fixed Sample About "t" Media Units Instead of One

By Jerome D. Greene
MarketMath, Inc.

A sample is interviewed on one unit of a media vehicle: let the variance-reciprocal or "information" of the average-unit audience estimate be indexed at "100". We now interview the same size sample on "t" units instead of one. What is the corresponding index of the amount of information?

Let p_i = person i's personal probability of exposure to one unit—his long-run proportion of units seen or heard.

Let $\hat{p}_i$ = the survey estimate of p_i obtained by dividing the number of units he saw/heard, r_i, by the number of units surveyed, t.

Let $\overline{p} = E(\hat{p}) = E(p)$ be the mean or average-unit audience proportion in the population. Then:

$$\begin{aligned}\sigma^2_{\hat{p}} &= E(\hat{p} - \bar{p})^2 \\ &= E\,[(\hat{p} - p) + (p - \bar{p})]^2 \\ &= E(\hat{p} - p)^2 + E(p - \bar{p})^2 = E(\hat{p} - p)^2 + \sigma^2_p \qquad (1)\end{aligned}$$

The two deviations are independent and thus their expected cross-product vanishes.

For person i:

$$\begin{aligned}E(\hat{p}_i - p_i)^2 &= E\left(\frac{r_i}{t} - p_i\right)^2 \\ &= \frac{1}{t^2} \bullet E\left(r_i - t \bullet p_i\right)^2\end{aligned}$$

This expectation term, for any one person, is the variance of the binomial frequency distribution of r_i trials given his personal probability p_i, and thus equals tp_iq_i. Therefore:

$$\begin{aligned}E(\hat{p}_i - p_i)^2 &= \frac{1}{t^2}\ (t \bullet p_i \bullet q_i) \\ &= \frac{p_iq_i}{t} \qquad (2)\end{aligned}$$

For all people, the expectation of equation (2) is:

$$E(\hat{p} - p)^2 = E\left(\frac{pq}{t}\right) = \frac{1}{t}\ (Ep - Ep^2)$$

$$= \frac{1}{t}\ [\bar{p} - (\sigma_p^2 + \bar{p}^2)]$$

$$= \frac{1}{t}\ (\bar{p} \cdot \bar{q} - \sigma_p^2) \qquad (3)$$

Therefore, substituting equation (3) into (1):

$$\sigma_{\hat{p}}^2 = \frac{\bar{p} \cdot \bar{q}}{t} - \frac{\sigma_p^2}{t} + \sigma_p^2$$

$$= \frac{\bar{p} \cdot \bar{q}}{t} + \left(\frac{t-1}{t}\right) \cdot \sigma_p^2$$

For an effective sample base of s, the variance of the average-unit audience estimate, $\bar{\hat{p}}$, is:

$$\sigma_{\bar{\hat{p}}}^2 = \frac{1}{s} \cdot \sigma_{\hat{p}}^2$$

$$= \frac{1}{s \cdot t}\ [\bar{p} \cdot \bar{q} + (t-1) \cdot \sigma_p^2] = \sigma_t^2 \qquad (4)$$

For convenience, let $\sigma^2_t = \sigma^2_{\bar{\hat{p}}}$ for a particular value of t.

The Statistical Efficiency (SE_t) of interviewing on t units, instead of 1, is the ratio of the reciprocals of the corresponding variances. Thus:

$$SE_t = \frac{\sigma^2_1}{\sigma^2_t}$$

$$= \frac{\dfrac{\bar{p} \cdot \bar{q}}{s}}{\dfrac{1}{s \cdot t}\,[\bar{p} \cdot \bar{q} + (t-1)\,\sigma^2_p]}$$

$$= \frac{t}{1 + (t-1)\left(\dfrac{\sigma^2_p}{\bar{p} \cdot \bar{q}}\right)} \qquad (5)$$

$\dfrac{\sigma^2_p}{\bar{p} \cdot \bar{q}}$ is constant for a particular media vehicle, and approximately constant for certain classes of vehicles. It is simply a variance ratio: the actual variance of the probability distribution divided by the maximum possible value of this variance. The actual variance ranges between zero (when all people have the same $p = \bar{p}$) and the maximum variance $\bar{p} \cdot \bar{q}$ (when all people have a p of either 0 or 1).

When all people have the same $p = \bar{p}$, equation (5) gives $SE_t = t$: the Statistical Efficiency increases directly with t, and interviewing with a fixed sample-size on t units provides t times more information than interviewing on one unit. When all people have a $p = 0$ or 1, $SE_t = 1$: the Statistical Efficiency is fixed at unity, and interviewing on t units provides no more information than interviewing on one unit.

$\dfrac{\sigma^2_p}{\overline{p} \cdot \overline{q}}$, the single parameter of equation (5), may be estimated from any audience survey covering two or more units of the media vehicle in question. This quantity is known to vary much more slowly in time than $\overline{p}$ itself, so that an old survey may well be used for planning a new survey.

For illustration, assume the personal probabilities p are Beta-distributed, in which case:

$$f(p) = \frac{1}{\beta\,(m, n)}\; p^{m-1} q^{n-1}$$

$$\frac{\sigma^2_p}{\overline{p} \cdot \overline{q}} = \frac{1}{m + n + 1} \qquad (6)$$

Substituting equation (6) into equation (5)

$$SE_t = \frac{t\,(m + n + 1)}{(m + n + t)} \qquad (7)^*$$

[*] Equation (7) is the special case of equation (5) when one assumes the Beta Distribution, and was first stated by Lester Frankel in 1968. (See Frankel's chapter, "The Role of Accuracy and Precision of Response in Sample Surveys", in *New Developments in Survey Sampling*, edited by Norman L. Johnson and Harry Smith, Jr., © 1969 by John Wiley & Sons, Inc., New York; or a reprint of this chapter available from Audits & Surveys, Inc., of which Mr. Frankel is Executive Vice President.) We know no prior statement of the general model given in equation (5) and in the two subsequent paragraphs concerning its limits.

Further illustrating, assume for a given media vehicle (or class of vehicles) that m + n = 1. Then question (7) reduces to:

$$SE_t = \frac{2 \cdot t}{(t + 1)} \tag{8}$$

For various values of t, equation (8) gives the following table:

t	$SE_t \times 100$
1	100
2	133
3	150
4	160
5	167
.	.
.	.
.	.
∞	200

For instance, in this particular example, interviewing a fixed sample about five media units instead of two adds 25% to the statistical information (167 ÷ 133 = 1.25).

In general, interviewing a fixed sample about t units instead of one reduces the *Variance* of the average-unit audience estimate by $(1 - 1/SE_t)$ and the *Standard Error* of estimate by $(1 - \sqrt{1/.SE_t})$. The following table gives the reductions in Variance and in Standard Error for our example assuming the Beta distribution with m + n = 1:

t	Reduction in:	
	Variance	Standard Error
	%	%
1	0	0
2	24.8	13.3
3	33.3	18.4
4	37.5	20.9
5	40.0	22.5
•	•	•
•	•	•
•	•	•
∞	50.0	29.3

Appendix

D. A Note About the Reliability of Cume Ratings vs. Average Ratings

By Jerome D. Greene
MarketMath, Inc.

1. Definitions

In broadcast audience research, a "Cume Rating" or "Cumulative Rating" estimates the percent of total people or households in a market exposed one or more times during a specific time period to a specific station or program broadcast. An "Average Rating" estimates the percent of total people or households in a market exposed during the average of two or more specific time periods to a specific station or broadcast, and may be derived by separately tabulating each specific audience and then averaging them together.

Thus, an Average Rating is the average of two or more Cume Ratings. For example, the percent of all television households in a market viewing station WAAA between 7:30 PM and 7:45 PM on Monday, January 22 is the Cume Rating of that specific station-time segment. Similarly, Cume Ratings are obtained for station WAAA between 7:30 PM and 7:45 PM on Tuesday, Wednesday, Thursday and Friday, January 23, 24, 25 and 26. The average of these five specific Cume Ratings is the Average Rating of station WAAA between 7:30 PM and 7:45 PM on *weekdays*, January 22-26.

Cume Rating		Station WAAA, 7:30 PM-7:45 PM
January 22		5.2
January 23		5.0
January 24		5.1
January 25		5.3
January 26		4.9
	Total	25.5
Average Weekday Quarter-Hour Rating		5.1

It is not often realized that the audience of one specific quarter-hour is a Cume Rating, but in fact this is so, because this conforms to the definition of exposure one or more times during a stated time period. Normally, one thinks of Cume Ratings for time periods comprising several or many quarter-hours, such as television's 7:30-11 PM "prime-evening-time" period.

In this case, Cume Ratings may be shown separately for each day or across a series of days. For instance, the percent of total television households viewing station WAAA at least once between 7:30 PM and 11 PM on Monday, January 22 is the Cume Rating for that station-daypart on that particular *day*. And the percent of total television households viewing at least once between 7:30 PM and 11 PM on at least one day from January 22 to 26 is the Cume Rating for that station-daypart during that entire *week* (five weekdays). Clearly, the *weekly* Cume Rating exceeds the *daily* Cume Rating to the degree that households view at least once during the week but not in every day thereof—in other words, to the degree that people do not have completely regular daily behavior patterns.

Given daily and weekly Cume Ratings, what kinds of Average Ratings go with them? The most common Average Rating is simply the average of all the specific quarter-hours which make up the time span of the Cume Rating. For instance, there are *14* quarter-hours in the 7:30-11 PM period on Monday, January 22, and the average of their audiences is the Average Rating for 7:30-11 PM on this particular *day*—as opposed to the Cume Rating for this time period of this day.

Across all five weekdays, moreover, there are *70* quarter-hours in this 7:30-11 PM period, and the average of their audiences is the Average Rating for 7:30-11 PM in this entire *week* (five weekdays)—as opposed to the Cume Rating for this time period during this entire five-day period.

Finally, there is a third kind of Average Rating, less often shown but often useful, called the "Average Cume Rating". As the name implies, this is the average across the week of Cume Ratings for two or more time segments within each day. For example:

Cume Rating		Station WAAA, 7:30-11 PM
January 22		8.8
January 23		8.7
January 24		8.8
January 25		8.9
January 26		8.8
	Total	44.0
Average Weekday Three-and-a-Half-Hour Cume Rating		8.8

2. Sampling Reliability

It is universally recognized by statisticians but largely unknown by media experts that Average Ratings are more stable than Cume Ratings, from the viewpoint of random sampling error. As explained before, an Average Rating is the average of Cume Ratings. The *Monday* 7:30-11 PM Average Rating of a television station is the average of its *14* component quarter-hour Cume Ratings; and *weekday* 7:30-11 PM Average Rating is the average of its *70* component quarter-hour Cume Ratings; and the *weekday* 7:30-11 PM Average Cume Rating is the average of its *five* component three-and-a-half hour Cume Ratings.

When things are averaged, stability is gained. When Cume Ratings are averaged to get Average Ratings, the Effective Sample Base or ESB is increased and the sampling error reduced. The gain in Effective Sample Base depends upon people's regularity of exposure over the time periods whose Cume Ratings are averaged to get the Average Rating: the more *regular* the exposure, the *less* the gain; the more *irregular* the exposure, the *greater* the gain.

For example, suppose Mr. Smith watched station WAAA on Monday evening January 22 in both the 7:30-7:45 PM and the 7:45-8 PM periods, while Mr. Jones did *not* watch station WAAA in either of these two quarter-hours. In this case, each person's behavior in the second period is the same as his behavior in the first. There is thus complete regularity of viewing across the two periods. There is no statistical gain from measuring the second period after the first, and therefore there is no increase in the Effective Sample Base by averaging these two people over the two time periods to get the Average. Quarter-Hour Rating for the Monday period 7:30-8 PM.

The degree of this "regularity" of exposure across averaged time periods is expressed by the statistician's measure of "correlation" on a scale from +1.0 through 0.0 to −1.0. A correlation of +1.0 expresses perfect regularity as illustrated; a correlation of 0.0 expresses complete independence or randomness of each person's behavior from one period to the next; and a correlation of −1.0 expresses complete irregularity which, in our example, would mean that if a person viewed in the first period then he did *not* view in the second, and vice-versa.

In the media field, negative correlations can hardly occur so we need concern ourselves only with correlations from 0.0 up to +1.0. With "random" behavior between averaged time periods—a correlation of 0.0—the Effective Sample Base for an Average Rating is equal to the sample size multiplied by the number of periods that are averaged. But with "completely regular" behavior between time periods—a correlation 1.0—there is no gain at all in Effective Sample Base for Average Ratings.

The true correlation lies between 0.0 and 1.0 and varies according to the number of averaged time periods, length of each period, time of day, media vehicle, and population group under analysis. Thus the Effective Sample Base for Average Ratings is always larger than the sample size, but the multiplier depends upon the particular situation. The more regular people's behavior across the averaged time periods, the closer this multiplier to 1.0 and the less the gain in ESB; the more random people's behavior across these time periods, the closer the multiplier to the number of periods being averaged and the greater the ESB gain.

The statistical theory and measurement of correlation are precisely defined in formulas, but the formulas merely quantify a simple common-sense proposition:

(1) If each person behaved exactly the same way in each time period, one time period would tell us all there is to know from our sample of "n" number of people; nothing would be gained by averaging two or more of these time periods, and the Effective Sample Base would simply be equal to "n".

(2) If on the other hand each person's behavior in each time period were completely independent or unpredictable from his behavior in any other time period, then each of the "t" number of averaged time periods would give us fresh new information from our sample of "n" people, each time period would be equivalent to a new sample of "n", and the ESB would equal "tn" (t × n).

(3) The truth, of course, lies between these two unreal extremes. The sheer fact that ratings vary somewhat among the averaged time periods proves that behavior is not completely regular and thus that the ESB for Average Ratings is greater than "n". On the other hand, the strong similarity of ratings between consecutive time periods of the same day, and equivalent time periods of different days, proves that there is considerable regularity and thus that the ESB for average ratings is much less than "tn".

3. How Reliable Are Cume Ratings and Average Ratings?

The American Research Bureau assigned to MarketMath the major job of determining the reliability of both Cume and Average Ratings, working from specific Arbitron Television survey data. Let us first review the method used*, and then discuss the resulting data:

(1) Compute directly the sampling error of each of many hundreds of published ratings from each of many different Arbitron Television samples, by dividing each sample into random parts and observing each rating's variability among these replicates.

(2) Comparê this actual sampling error with the hypothetical error if ESB = n—that is, if the Effective Sample Base were simply equal to the sample size. The ratio of "hypothetical" to "actual" sampling error is defined as the "Statistical Efficiency" or "SE" of the sample, for that specific rating.

(3) If the SE is *less* than 1.0, the ESB is *less* than the sample size n. If the SE is *greater* than 1.0, the ESB is *greater* than n. In general, ESB = SE × n: "Effective Sample Base Equals Statistical Efficiency times Sample Size".

(4) This was done separately for hundreds of Cume and Average Ratings of specific stations and programs, by time period and by demographic group within the total sample.

Computing the sampling error of each rating from its variability among random parts of the total sample avoids any assumptions about the correlations or regularity of exposure

*A comprehensive discussion of procedures used in the study is presented in Chapter II of this report.

among time periods, or about the effects of sample clustering, stratification, and weighting to match population demographics. All these factors which influence sampling error, over and above the raw sample size n, are automatically taken into account.

If the rating is highly variable among people or households, it has a large sampling error by definition, and this is automatically revealed by its high variability among the sample replicates—each of which is a separate, independent mini-sample of the population (households, total people, or demographic group). Conversely, if the rating varies only slightly among people or households, it has a small sampling error, and this is automatically revealed by its low variability among the sample replicates.

In this way we derive the actual Standard Error ("sigma") and Variance ("sigma-squared") of each rating. Now we compare this actual variance with the hypothetical variance if ESB = n, using the familiar formula pq/n where p = the rating as a percent, q = 100 − p, and n = sample size. The ratio of "hypothetical" to "actual" variance is SE: "Statistical Efficiency".

Finally, we compute the Effective Sample Base (ESB = SE × n), which may be larger or smaller than n. The utility of ESB is that we obtain the actual, correct Variance if we substitute ESB for n in the familiar formula—giving pq/ESB. The square-root of this is the important Standard Error as a plus-minus sampling error margin in percentage points around the reported rating p.

Different surveys have different sample sizes (n), so it is convenient to analyze the Statistical Efficiencies—the multipliers of n to get the Effective Sample Bases. This we have done for thousands of specific Arbitron Television ratings, and a clear pattern emerges. Over stations/programs, days, and times-of-day, the following table shows the relationship of the Statistical Efficiency (SE) to the number of averaged time periods (t) for household ratings:

Number of Averaged Time Periods (t)	Approximate Statistical Efficiency (SE)
1	0.9
2	0.9
5	1.1
10	1.4
18	1.8
30	2.3
32	2.4
40	2.8
50	3.2
60	3.6
70	4.0
84	4.4
90	4.6
98	4.9
420	10.7

Note that Cume Ratings involve no averaging; therefore "t" equals 1, and the Statistical Efficiency of Cume Ratings is 0.9 as shown in the table for $t = 1$. To obtain the Effective Sample Base (ESB) of any Cume Rating, simply multiply the sample size (n) by 0.9. Then the Standard Error is approximately $\sqrt{pq/ESB}$ where "p" is the rating as a percentage ($q = 100 - p$).

Now consider an Average Quarter-Hour Rating based on the average of ten specific quarter-hours (such as Monday, January 22, 4-6:30 PM). The table shows that the SE = 1.4 when t = 10. Multiply the sample size by 1.4 to get the ESB, and then calculate $\sqrt{pq/ESB}$ to get the approximate Standard Error of this rating.

Most reported Average Quarter-Hour Ratings are averaged across an entire week (five-day or seven-day), with a very large number of quarter-hours entering into the average. For instance, the Average Quarter-Hour Rating 4-6:30 PM, Monday-Friday, involves 50 quarter-hours. The table shows that the SE = 3.2 when t = 50. So multiply the same sample size by 3.2 to get the ESB, and then calculate $\sqrt{pq/ESB}$ to get the approximate Standard Error of the rating.

4. Conclusions

The key conclusions of this paper are:

(1) Average Ratings are more reliable—i.e., have smaller sampling errors—than Cume Ratings.

(2) The Effective Sample Base of Average Ratings increases, but at a decreasing rate, with the number of specific time periods (e.g., quarter-hours) which are included in the average.

(3) The increase in Effective Sample Base of Average *Weekly* Quarter-Hour Ratings is substantial because of the large number of quarter-hours which enter into the average.

Appendix

E. Glossary of Terms Used in Arbitron Television Reports

Area of Dominant Influence (ADI)—The Area of Dominant Influence is a geographic market design which defines each market exclusive of another based on *Measurable Viewing Patterns.* As the name implies, the ADI is an area that consists of all counties in which the home market stations receive a preponderance of viewing. Each county in the U. S. (excluding Alaska) is allocated exclusively to only one ADI*. There is no overlap.

The original ADI allocations were based on a 1965 county-by-county study of television circulation using the viewing data obtained by diary from approximately 250,000 television households. From these viewing data, Arbitron prepared estimates of the total viewing hours in each county for an average week, and the percentage of the estimated total viewing hours of each station for which viewing was reported. The original ADI allocations were based on these figures.

The ground rules for ADI allocations are relatively simple. Once the estimated total viewing hours for a county, and the percentage of such estimated total for each station, are known, Arbitron sums the station percentages by market of origin. The market of origin having the largest total percentage is deemed to be the "dominant influence" in the county under consideration, and that county is allocated for ADI purposes to that market of origin. An additional analysis, based on

*Where a county is divided by Arbitron into more than one sampling unit, each unit is analyzed as if it were a county for ADI purposes, and is assigned to an ADI on the basis of the rules described above.

share of viewing hours in fringe time periods, is also performed in some cases.

There are exceptions to the general rule above:

(A) Arbitron reserves the right to exercise its judgment in the case of counties with unusual physical features or peculiar marketing considerations.

(B) If its home station achieves at least a 20 share, a Metro county, or the Home County of a station having no Metro Rating Area, or the Home County of an S-2 satellite station, is not assigned to the ADI of another market *unless* the average of the percentages of viewing hours of the stations in the other market is at least 10% greater than the sum of the percentages of the viewing hours of the stations in the Metro or Home County under consideration.

(C) To re-assign a county from one ADI market to another, a minimum of 15 in-tab households is required.

(D) In considering the creation of a new ADI market, the criteria for the assignment of counties to an ADI would prevail; in addition, a market must win its Home County, and that Home County must have at least 10,000 television households.

Adjacent Areas of Dominant Influence (Adjacent ADI's)— Viewing is reported in a maximum of three adjacent ADI's served by Home Market stations. These adjacent ADI's lie within the Home Market's TSA, but outside of the Home Market's ADI. Where more than three adjacent ADI's lie within a market's TSA, selection of the three to be reported is based on an analysis of the TV household contribution to each adjacent ADI and other pertinent viewing characteristics. The ADI's to which counties in the TSA have been assigned are identified by codes which appear above the county listing on Page 5 of the report. Counties with the code "O" lie within the ADI of a market other than the three adjacent ADI's reported. The TV households totals of adjacent ADI markets are also reported.

Average Quarter-Hour Audience
(See "Quarter-Hour Audience")

Color Set Penetration—Arbitron reports estimates of color TV households penetration for the TSA, the ADI and Metro of all Metro markets; the TSA and ADI of all non-Metro markets; and the TSA of all non-ADI markets. These estimates are based on information obtained during the diary placement interview.

Controls—Arbitron weighting techniques are used in all sampling units to establish proportionate representation of viewing by Age of Head-of-Household and by week. The weighting techniques are also used in certain sampling units containing CATV households, and in certain sampling units where special interviewing techniques are used.

Cume Households—An estimate of the number of different television households that viewed each reported station at least once during the average week for five continuous minutes or more during the reported time period. This is also called the cumulative or unduplicated audience, or circulation. Estimates are based on viewing in the Total Survey Area only.

Cume Persons—An estimate of the number of different persons who viewed each reported station at least once during the average week for a period of five continuous minutes or more during the reported time period. Estimates are based on viewing in the Total Survey Area only. (See also "Cume Households.")

Demographic Rating—Viewing estimates of persons in a particular sex-age group divided by the total number of persons in television households in that category. The result is rounded and expressed as a whole percentage or rating. The Audience Category Chart shows which demographic categories are reported in each report section.

Effective Sample Base (ESB)—The sample size to be used for assessing the statistical variance of audience estimates.

HPDV
Households-per-Diary Value.

HPRP
Households per ADI Rating Point.

Housewife
The female head-of-household age 16+.

Home County
See "Metro Rating" below.

Households Using Television (HUT)—An estimate of the number of unduplicated households (with one or more sets tuned in) which viewed all television stations during the average quarter hour of the time period. HUT is expressed as a percentage of the total number of television households in the Metro, ADI or Home County.

In-Tab Sample—The number of television households which returned diaries tabulated in the production of the report.

Metro (or Home County) Rating Area—Metro Rating Areas, where applicable, generally correspond to Standard Metropolitan Statistical Areas as defined by the U. S. government's Office of Management and Budget, subject to exceptions dictated by historical industry usage and other marketing considerations such as channel allocations. (Home Market MRA counties are indicated in the listing on Page 5 of the report by an "M" preceding the county name.)

Where there is no defined ADI, ratings may be shown for the Home County of the station's city of license. The Home County is indicated in the listing by an "H" preceding the county name on Page 5 of the report.

Multi-Set Penetration—Arbitron reports penetration estimates of households with more than one television set in the TSA, the ADI and Metro of all Metro markets; the TSA and ADI of all non-Metro markets; and the TSA of all non-ADI markets. These estimates are based on information obtained during the diary placement interview.

Net Weekly Circulation—The estimate of the number of unduplicated households and the number of unduplicated adult persons which viewed a station at least once a week for a period of five continuous minutes or more.

Original Sample Size—The number of television households originally drawn for the survey.

PVT (Persons Viewing Television)—In the ADI, the total number of persons viewing all television is reported as an

ADI rating on the HUT/PVT/TOT line for each time period. This estimate includes viewing to both reported and non-reported stations (those stations whose audiences were too small to meet minimum reporting standards).

Projection—The expansion of sample statistics to population or households information in the respective universe.

Quarter-Hour Audience—A projected estimate of the unduplicated audience having viewed a station for a minimum of five continuous minutes within a specific quarter hour. These quarter-hour total audiences, when combined in time, become Average Quarter-Hour Audiences.

Rating—The estimated number of television households (or persons in a particular sex-age category) viewing a station for at least five continuous minutes during an average quarter hour of the reported time period, expressed as a percentage of all television households (or persons in the sex-age category) in the reported area. When the rating is estimated to be less than 0.5% for a time period the space is left blank; this blank is not intended to imply that no viewing occurred.

Sampling Unit—A sampling unit normally is one county, although some counties have been divided into two or more sampling units because of population distribution, terrain or special interviewing technique areas.

Satellite Station—A station that duplicates some or all of the programming of a parent station in order to serve an area not normally reached by the parent, and which is assigned separate call letters and channel number by the FCC.

Share—The percentage of the total Households Using Television (HUT) reached by a station during the specified time period.

TOT—Total TSA viewing.

Total Survey Area (TSA)—A geographic area comprising those counties in which an estimated 98% of the net weekly circulation of commercial home market stations occurs. Estimates of viewing in the Total Survey Area are reported in thousands.

Universe—All television households located in the specified area.

For additional information the reader is directed to "Standard Definitions of Broadcast Research Terms," published by the National Association of Broadcasters, 1771 N Street, N.W., Washington, D.C. 20036.