# Stratification for Radio Listening Estimation

Richard Griffiths
Arbitron, Inc.

## Abstract

To form weights for the sample respondents to its quarterly surveys, Arbitron uses a raking (or, sample balancing) procedure. For years, there have been three marginal variables in this raking procedure. Recently, a fourth variable was added to the mix. This paper examines issues related to adding the new marginal variable. In particular, we examine different stratification configurations of the new variable and their effect on the statistical properties (e.g., mean square error) of radio listening estimators. Additionally, the population controls for this new variable are based on a survey with relatively small sample sizes. We examine the effect of these stochastic population controls on the precision of the estimators.

**Keywords**: Stratification, Raking, Mean square error, Stochastic population controls

## Introduction

What I'll present in this paper is a case study in stratification in the media research industry. In the paper, I examine some of the issues – both statistical and business-oriented – involved in setting the appropriate level of stratification (post-stratification) to apply to a repeated survey of radio listening behavior in the United States.

From the statistical perspective, the issues boil down to how to balance the tradeoff between increased variance due to smaller weighting cell sample sizes, which come from additional levels of stratification, versus reduced potential for bias due to finer levels of stratification, which limit the potential for non-response and coverage bias. This plays out in a series of statistical analyses – some regression analyses to look at potential reduction in bias, examination of variation in weights under different stratification levels and its effect on the precision of the final estimators, and, finally, in terms of mean square error.

From the business perspective, there is customer perception that increased levels of stratification are good. In fact, there is the perception that "more stratification is likely to make the ratings of my radio station go up." This customer perception obviously has some pull on the internal business decision and I examine its effect on the decision-making process, both in the present and in the future.

Towards the end of the paper, and as something of a sidebar, I look at an issue related to the population controls used in the survey weighting procedures. The survey weighting process involves raking the sample to a set of population controls, some of which are stochastic (i.e., they are subject to sampling error). I examine the effect of this source of randomness on the precision of the final estimators. Examination of this issue was not part of the original business decision on the level of stratification and I discuss its likely effect on the future decisions to be made.

## Background on the Survey and Its Methodology

To produce estimates of radio listening audiences in the United States, Arbitron divides the country into about 300 geographical areas called markets. Arbitron then conducts surveys of an RDD sample in each market. Each survey is conducted over a 12-week period. About 100 of the markets are surveyed four times per year; the others are surveyed two times each year.

To ensure the selected sample represents the demographic and geographic characteristics in each market, Arbitron uses a raking methodology to weight the sample to the population. Because this is the only stage of weighting employed in Arbitron's surveys, the raking compensates for non-response and non-coverage. It thus serves to reduce the bias from non-response and non-coverage. (See Kalton and Flores-Cervantes, 2003.)

Using the weights from the raking procedure, expansion estimators of the numbers of people who listen to the various radio stations and the amount of time people spend listening, among other radio listening behaviors, are then constructed.

Traditionally, Arbitron has used a combination of age and gender, race/ethnicity, and geography (e.g. county) as marginal variables in the raking procedures in each market. Beginning in 2006, Arbitron has added a new marginal variable – primary language for the Hispanic population. This new variable will be added to the

raking procedures in 21 of the markets Arbitron measures.

Inside Arbitron, we colloquially refer to the addition of this variable as language weighting and I will do similarly throughout this paper.

## Statement of the Problem

Weighting represents a tradeoff. On the one hand, the addition of weighting variables and classes will generally reduce the potential for (non-response and non-coverage) bias in survey estimators. On the other hand, it will tend to increase the variance of those estimators. The question I seek to answer in this paper is, what level of language weighting offers the best tradeoff – the greatest reduction in the potential for bias and the least reduction in precision – for Arbitron's radio listening estimates?

This is primarily a statistical question, but it warrants an answer that addresses the concerns of Arbitron's clients and that can be understood by them. The clients have a large stake in methodological changes such as this and need to be informed of the reasons for the decisions we make. Thus, this very statistical question and its answer need to be communicated to non-statistical clients, both internally and externally. Our analyses are designed with this in mind.

My formulation of the answer to the question began with a set of possible stratifications. At one extreme, Arbitron could use a very fine Spanish language stratification with four language classes – All Spanish, Mostly Spanish, Mostly English, and All English – with Arbitron's traditional 16 age/gender classes for each language class. This would result in 64 overall classes for language weighting and would give the greatest reduction in potential bias. But, because the sample would be spread thinly over this many classes, the variation in sampling weights would be very large. This would result in the greatest loss of precision among the possible stratifications.

At the other extreme, Arbitron could use a very broad Spanish language stratification with two language classes – Spanish Primary and English Primary – and no age/gender classes. This would result in two overall classes for language weighting and would give the least impact on current precision. It would also result in the least reduction in potential for bias. In between, there are any number of combinations of the four language and 16 age/gender classes.

So, from the outset, Arbitron made a concession to client concerns – there is not a no-language weighting option in our set of possible stratifications. It's presumed there will be some form of language weighting; the question is how much.

This paper reports the results of an empirical analysis designed to determine the optimal level of language stratification in the face of the stated tradeoff.

## The Empirical Study

What level of language weighting offers the best tradeoff between bias and variance? To answer this question, I conducted an empirical study. This study consisted of several components, some client-oriented and some not so client-oriented. The client-oriented analyses were

- a regression analysis to give a simple measure of the potential for reduced bias under different levels of stratification; and
- an analysis of the variability of weights to examine the relationship of the different levels of stratification to variance.

Both of these analyses allowed me to pull out some concepts that were familiar to most clients, or at least that were more readily explainable to clients. They also offered some nice opportunities to tell the story graphically, which is helpful when communicating with non-statistician clients.

The primary statistical analysis was an analysis of the estimated MSE of radio ratings estimators under various stratification schemes; this analysis encapsulated the variance/bias tradeoff.

These analyses were intended to support a business decision on language weighting and also to serve as the technical background for delivering the rationale for our decision to clients. So, while from a statistical perspective, the MSE analysis was the primary analysis, the other analyses allowed me as a statistician to communicate the message in a simpler manner, using statistical techniques more readily understood by the layman.

## Regression Analysis Methodology

To communicate the idea of reduced bias via stratification, I turned to multiple regression (ANOVA) modeling. The form of the model was

$$y_i = \beta_0 + \beta_{11}x_{11i} + \beta_{12}x_{12i} + \ldots + \beta_{1c_1}x_{1c_1 i} + \beta_{21}x_{21i} + \ldots + \beta_{2c_2}x_{2c_2 i} + \beta_{31}x_{31i} + \ldots + \beta_{3c_3}x_{3c_3 i} + \beta_{41}x_{41i} + \ldots + \beta_{4c_4}x_{4c_1 i} + e_i$$

,where $y_i$ is a measure of the amount of listening for person $i$ to Spanish-format radio; $x_{vji}$ (assumed fixed) is 1 if person $i$ is in class $j$ of weighting variable $v$ and 0 otherwise; and $e_i$ is the random error term.

There were four sets of independent (dummy) variables in the model. These corresponded to the four weighting variables – age/gender, geography, race/ethnicity, and language. The $c_j$ indices represented the number of marginal classes in each weighting variable.

This model allowed me to examine how well variation in radio listening behavior was explained by the weighting variables. By varying the classes of the language variable in the model, I could look at how well different language stratification schemes explained variation. This concept could be communicated to internal staff and customers.

In all, I studied 40 different language stratification schemes that pretty much ran the gamut between the two extremes, along with a no-language stratification, which represented the current Arbitron stratification scheme. I performed regression analyses separately for each of the 21 language-weighting markets.

The main reason for turning to regression analysis to look at the potential for reduction of bias under each stratification scheme was that the $R^2$ value provides a nice summary of the results of the modeling for each stratification scheme and was generally accessible to customers and internal non-statisticians, much more so than is an estimated bias value.

To communicate the results of the regression modeling, I formed two benchmarks. The first benchmark was the $R^2$ value for the no-language weighting stratification (the base). The second was the $R^2$ value for the infeasible stratification (the peak). In communications with internal non-statistician staff and customers, the focus was on how much a given language stratification improved the $R^2$ value over the base $R^2$ value (no language weighting), as well as how close the $R^2$ value for a given language stratification was to the peak $R^2$ from the infeasible stratification. As a general rule, I told them that we want a language stratification that gives an $R^2$ value noticeably larger than the base $R^2$ value; otherwise, the language stratification may not provide a reduction in bias. Additionally, the closer the $R^2$ value is to the peak $R^2$ value, the better – as I told them, the better the stratification is doing at grouping people with similar listening characteristics. That is, if we consider the difference between the peak $R^2$ value and the base $R^2$ value to be the maximum achievable gain, the larger the proportion of the maximum achievable gain attained by a stratification, the better.

## Weight Variability Analysis Methodology

To communicate the idea of increased variance from finer stratification, I turned to a quantity called *statistical efficiency*.[1] This measure is used in the media research community for assessing the loss of precision that results from weighting.

The statistical efficiency $S$ of a sample is a function of the relative variance of the weights:

$$S = \frac{1}{1+L},$$

where $L$ is the relative variance of the weights. (See Kish, 1992.)

The effective sample size $ESS$ is a function of the statistical efficiency:

$$ESS = S \cdot n,$$

where $n$ is the sample size.

While the increase in variance of adding stratification classes was the measure of primary importance to me, mainly as a component of the MSE, graphs of the effect of increased weighting on statistical efficiency were very useful for communicating this effect internally and to customers.

## MSE Estimation Methodology

Mean square error, as the sum of the variance and bias squared, was the primary technical measure that I needed to communicate to support the decision. MSE encapsulates the tradeoff between reduction in bias and increased variance that comes from adding levels of post-stratification.

The idea of MSE was difficult to communicate to internal staff and customers. However, putting it in terms of a tradeoff between $R^2$ value and statistical efficiency helped. Though, it had to be noted that these were only surrogate measures for the variance and bias.

[1] *Statistical efficiency* is also known as *weighting efficiency* and is related to *design effect*.

To calculate variance estimates, I used jackknife methodology. Since Arbitron's surveys are RDD surveys with all people 12 years and older in a household selected for the sample, the jackknife technique was applied at the household-level. The raking and estimation procedures were run for each replicate to generate replicate estimates. Variance estimates were calculated from these.

As for bias, without knowing true population values for the radio listening estimators, it is difficult to establish a robust measure. Instead, in this study, I relied on a relative measure of bias.

To estimate the bias, I made the assumption that the finer the stratification, the lower the bias. The idea here is that if we have different response rates and coverage rates across different subgroups, the more we stratify, the more we tend to group individuals with similar listening behavior and similar response and coverage rates, or at least the less we have groups with disparate rates. Following this idea through, I assumed that the estimates generated under finest language classification had the least bias. I then measured the bias of the other stratifications relative to this by comparing their estimates to those of the finest language stratification.[2] To filter out some of the sampling variance in this bias estimate, I took an average over several surveys.

Take the following as an example of this approach: Suppose we have two language stratifications, with LC1 being finer than LC2. Suppose also that we calculate estimates for a particular characteristic under the two language classifications for four surveys. We then have $\hat{Y}_{1W}$, $\hat{Y}_{1Sp}$, $\hat{Y}_{1Su}$, and $\hat{Y}_{1F}$ as estimates under LC1 for the Winter, Spring, Summer, and Fall surveys. Similarly, we have $\hat{Y}_{2W}$, $\hat{Y}_{2Sp}$, $\hat{Y}_{2Su}$, and $\hat{Y}_{2F}$ under LC2. We are interested only in the relative bias of LC2 versus LC1, so we can assume LC1 has no bias. I.e., $E(\hat{Y}_{1j}) = Y_j$, for *j*=*W*, *Sp*, *Su*, and *F* and $Y_j$ is the true value of the estimand.

Then, $\frac{1}{4}\sum_j (\hat{Y}_{2j} - \hat{Y}_{1j}) = \frac{1}{4}\sum_j \hat{Y}_{2j} - \frac{1}{4}\sum_j \hat{Y}_{1j}$ is an estimate of the average bias of LC2 relative to LC1, since

$$\frac{1}{4}\sum_j E(\hat{Y}_{2j} - \hat{Y}_{1j}) = \frac{1}{4}\sum_j (E(\hat{Y}_{2j}) - Y_j) = \frac{1}{4}\sum_j Bias(\hat{Y}_{2j}) \cdot$$

## Empirical Study Results

For some of the more technical clients, we showed them graphs like the one given in Figure 1. This graph illustrated the tradeoff between the $R^2$ value from the regression analyses and the statistical efficiency and how that related to MSE.

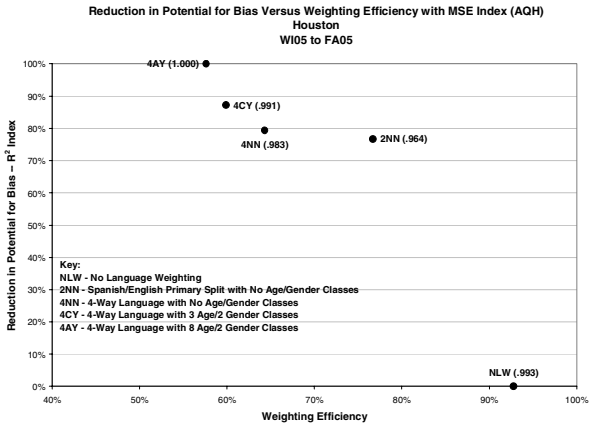


*Figure 1.* Reduction in Potential Bias v. Weighting Efficiency

For the less technical clients, we showed graphs that displayed the bias/variance tradeoff in terms of $R^2$ and statistical efficiency, without any initial reference to MSE. Examples of these graphs are given in Figures 2 and 3.

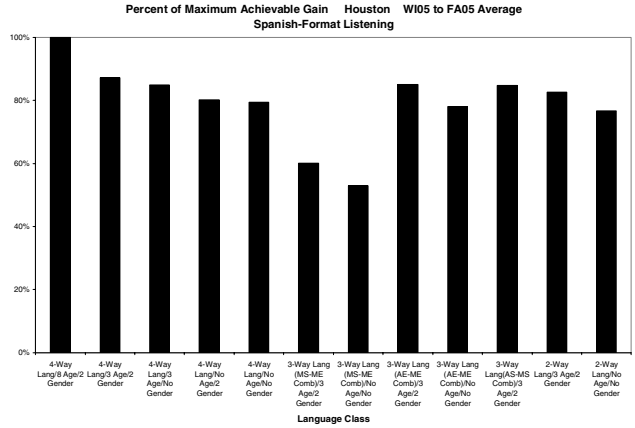


*Figure 2.* Percent of Maximum Achievable Gain in $R^2$.

[2] The finest stratification is assumed to have zero bias under this relative measure of MSE. This is equivalent to assuming that responses are missing at random within the subgroups of this stratification.

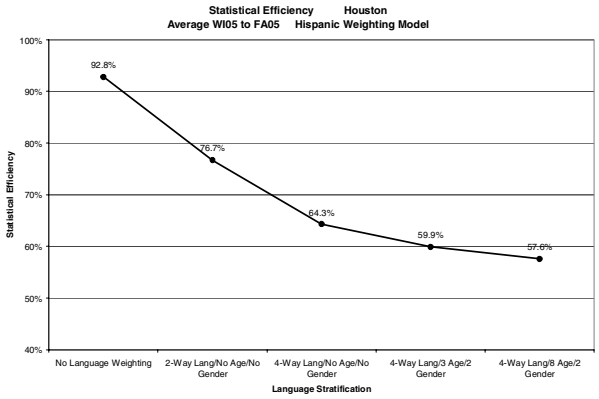


*Figure 3.* Statistical Efficiency.

The graphs like those in Figure 2 showed how $R^2$ value varied by stratification scheme. In general, we found that the broad Spanish language stratification with two language classes – Spanish Primary and English Primary – and no age/gender classes achieved a large proportion of the maximum achievable gain in $R^2$ value. This was true in most of the 21 markets.

Graphs like those in Figure 3 showed how statistical efficiency declines – because weight variability increases – as the number of strata increases. The decline was more precipitous in some markets than in others.

After showing clients graphs like those in the above figures to form a knowledge base for the bias/variance tradeoff, we were ready to turn to MSE results. In particular, we turned to graphs like the examples given in Figures 4 and 5.

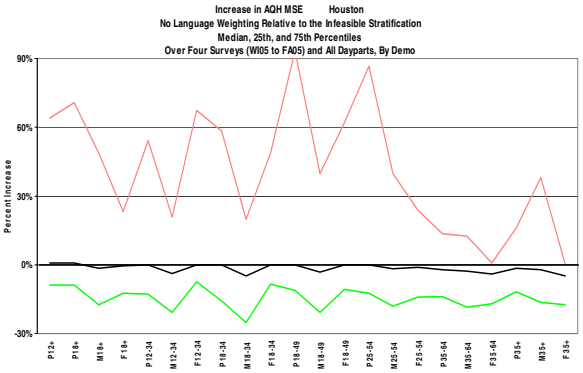


*Figure 4.* Increase in MSE: No Language Weighting.

These graphs give the median (black line) and 25th (green line) and 75th (red line) percentiles of the distribution of estimated MSE values over all radio rating estimators in the study for a particular market by demographic group.

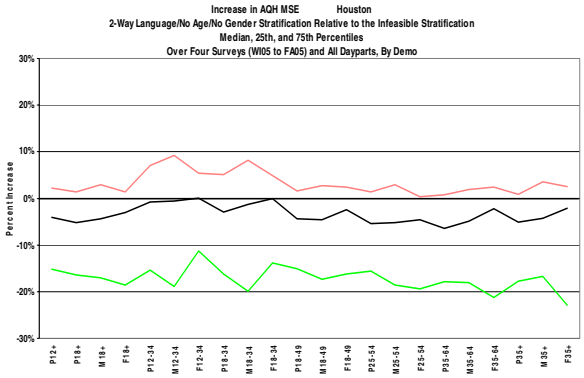


*Figure 5.* Increase in MSE: 2-Way Language Weighting.

Graphs like these formed the basis for our final decision on the optimal level of stratification. In general, the MSE results indicated that the two-strata language classification (Spanish/English Primary) gave the best distribution on MSE values. Again, this was clearer in some markets than in others.

## The Final Decision

Although both statistical and business issues factored into the final decision, it was ultimately based largely on the preceding statistical analyses. Both the internal and external clients accepted the results of the study – to use the two-way language stratification in all 21 markets – although some did so begrudgingly.

Even with the final decision made, though, there are still client concerns that weigh on the future of language weighting. One of the concessions we made to clients who begrudgingly accepted our decision was that we would continue to re-examine the data and our decision as we gain more experience with language weighting.

In re-examining our decision, one of the issues that will be factored in is the sampling error in the language population controls. The analyses leading to the final decision ignored the effect of stochastic population controls on the MSE. A preliminary study of this effect is given below.

## Afterword to the Final Decision

From a statistical perspective, the objective function used in making the decision is the MSE, which is expressed as the variance plus the bias squared:

$$MSE = V + B^2 .$$

This is the quantity that guided us through the original analyses and supported the final decision. However, as mentioned, not all clients were happy with our decision and we have agreed to continue to evaluate it as time progresses.

If we juxtapose this sentiment with the MSE concept, we get a quantity that not all of us are familiar with – business squared error:

$$BSE = V \quad + B^2 + C^2 ,$$

where *C* is a measure of client disapproval of our original level of language weighting.

I think in our original decision, the *C* term was near zero. However, I believe that as time progresses and as we continue to re-evaluate our decision with more analyses, the value of *C* is likely to grow. Then my job will become one of ensuring that the *V* and the *B* in the equation aren't ignored (i.e., capriciously set to 0 via some business decision).

## Stochastic Population Controls

### Background

The population controls we used in raking the sample to the language marginal were stochastic. They were obtained through sample surveys of the Hispanic population in each of the 21 markets. These surveys were carried out by a vendor under contract to Arbitron. The vendor created the population estimates under an agreed-upon methodology.

In our original analysis, we were unable to account for the effect of the stochastic population controls on the MSE values since the vendor did not provide us with any information on the distributional characteristics of the population estimates. (Some rough calculations were made under various assumptions, but I didn't feel comfortable with the assumptions underlying these calculations, so they were never included as part of the analysis.) Recently, however, we obtained information on their precision. In this section, under some simplifying assumptions, we investigate how the stochastic population controls affect the MSE of estimators under different stratification schemes.

### Methodology

To estimate the effect of the stochastic population controls on the MSE estimates, we conducted an empirical study in which we generated replicate estimates of the population controls and used them in a replicate variance estimation procedure for the radio listening estimates. The replicate population controls were generated using estimated standard error provided by the vendor.

The vendor did not supply estimates of the covariances of the population controls, so we were limited to generating replicate population controls under univariate distributional assumptions. We assumed the following:

$\hat{X}_{ir} \sim N(X_{i0}, S_i^2)$, $i$=1,2,3,4 (for the 4 language classes with no age/gender strata)

where $r$ indexes the replicate; $\hat{X}_{ir}$ denotes the estimated population estimate for the i$^{th}$ language class, $r^{th}$ replicate; $X_{i0}$ is the full-sample population estimate for class i; and $S_i$ is the estimated standard error of $X_{i0}$.

The replicate estimates were then generated from this distribution under the constraint that

$$\sum_{i=1}^{4} \hat{X}_{ir} = \sum_{i=1}^{4} X_{i0} .$$

We conducted the empirical investigation in one market – Los Angeles – for one survey. In particular, we were interested in the effect of stochastic population controls for 2-way versus 4-way language stratification with no age/gender classes.

For this market, we calculated variance estimates for the same radio stations and demographic subgroups and dayparts of the original language weighting study. But, this time, we used the replicate controls totals in producing the replicate radio listening estimates. We then compared these variance estimates to those generated under the assumption of fixed population controls (the assumption made in the original study). This gave us a measure of the increase in variance associated with stochastic population controls.

### Results

Table 1 gives some distributional measures for the increase in variance over no language weighting for the 2- and 4-way language stratifications with no age/gender strata. These summary measures are of the increases in variances over all radio stations and demographic subgroups and dayparts studied in the Los Angeles market for the fall 2005 survey.

**Table 1**
Distribution of Increase in Variance Over All Demos and Dayparts

| | Increase in Variance | | | | |
|---|---|---|---|---|---|
| **Language Stratification** | **25th Percentile** | **Median** | **Mean** | **75th Percentile** | **95th Percentile** |
| **4-way** | 0.0% | 2.4% | 5.8% | 8.2% | 27.6% |
| **2-way** | -0.2% | 0.7% | 3.9% | 4.5% | 22.7% |

As expected, the 4-way stratification leads to larger increases in variance than does the 2-way stratification.

This analysis and its results have at least a couple of ramifications:

- First, the effect of stochastic population controls needs to be included in our estimates of the precision of radio listening estimates. The effect of using stochastic population controls is not small enough to ignore.
- Second, it needs to be included in our MSE analyses as we re-evaluate the appropriate level of language weighting in the future.

## Discussion

In this paper, I presented the results of an empirical study aimed at supporting a business decision on what level of stratification to use in a media research survey. Some of the lessons learned from this study that can be carried into future studies are as follows:

- For presenting results to clients:
  - Clients were generally capable of understanding the basic ideas of the bias/variance tradeoff. It's important to capitalize on this fundamental understanding when explaining results.
  - Clients are capable of understanding the relationship of $R^2$ values to the potential for bias. In particular, they understand the concept that grouping people via post-stratification can reduce bias.
  - Clients are capable of understanding the effect of extra weighting on precision.
  - Graphs are good.
- What not to do when presenting results to clients.
  - Clients have a hard time understanding the concept of MSE, even when it is connected to more intuitive concepts.
    - Do not rely solely on an MSE analysis even though the MSE analysis supplies the primary foundation for decision making.
  - Do not present just the part of the results you think the clients will understand. (You can fill in the blanks on that one!)
- We need to incorporate the sampling error due to stochastic population controls in the analyses.

## References

Kalton, G. and I. Flores-Cervantes (2003), "Weighting Methods," Journal of Official Statistics, 81-97.

Kish, L. (1992), "Weighting for Unequal $P_i$," Journal of Official Statistics, 183-200.