

## An Examination Of Poststratification Techniques For The Behavioral Risk Factor Surveillance System

Michael P. Battaglia<sup>1</sup>, Martin R. Frankel<sup>2</sup>, and Michael Link<sup>3</sup>  
 Abt Associates Inc.<sup>1</sup>

Abt Associates Inc. and Baruch College, CUNY<sup>2</sup>

Centers for Disease Control and Prevention<sup>3</sup>

### Abstract

Random-digit-dialing surveys such as the Behavioral Risk Factor Surveillance System (BRFSS) typically poststratify on age by gender by race/ethnicity cells using control totals from an appropriate source such as the 2000 Census, the Current Population Survey, or the American Community Survey. Using logistic regression and CHAID we identified key “main effect” socio-demographic variables and important two-factor interactions associated with several health risk factor outcomes measured in the BRFSS. A procedure was developed to construct control totals, which were consistent with estimates of age, gender, and race/ethnicity obtained from a commercial source and distributions of other demographic variables from the Current Population Survey. Raking was used to incorporate main effects and two-factor interaction margins into the weighting of the BRFSS survey data. The resulting risk factor estimates were then compared with those based on the current BRFSS weighting methodology and mean squared error estimates were developed.

**Keywords:** Control Totals, Poststratification, Raking, Random-Digit Dialing

### 1. Introduction

Survey researchers are increasingly concerned about potential bias in random-digit dialed (RDD) surveys resulting from frame noncoverage and unit nonresponse. Households with no landline telephones, including those with only cellular telephones, are excluded from the RDD sample frame. Unit nonresponse is an issue in any of the various survey modes, but response rates to RDD surveys have been declining steadily (Curtin et al. 2005, Battaglia et al. 2006), in part because of growth in screening technologies, privacy concerns, telemarketing, and refusals. In this research, we attempted to reduce the potential for nonresponse bias in a major health survey by identifying and assessing changes in standard weighting procedures. The research shows that the addition of a few key variables to the weighting methodology can significantly reduce the potential for nonresponse bias.

### 2. Previous Research Examining Factors Related To Nonresponse

Rao et al. (2005) evaluated the degree to which noncoverage and unit nonresponse contributes to under-representation of important socio-demographic subgroups in RDD surveys. The Behavioral Risk Factor Surveillance System (BRFSS) -- a monthly RDD survey administered by all the states with assistance from the Centers for Disease Control and Prevention (CDC) to collect health-related information -- was used in the analysis. BRFSS is an important survey that generates state and local prevalence estimates among adults of the major health conditions and behavioral risks associated with premature morbidity and mortality. Rao et al. evaluated noncoverage and nonresponse in six states (California, Illinois, North Carolina, New Jersey, Texas, and Washington). Five of these states had experienced state-level response rates at or below 40% over the past several years, with North Carolina being the exception. The researchers compared the distributions of socio-demographic variables for these six states from the 2003 BRFSS with the distribution of the same variables from the March 2003 Current Population Survey (CPS). They found that the youngest age group (18-24 years) was highly under-represented in North Carolina, New Jersey, Texas, and Washington. In California and Illinois, it was under-represented but not substantially. Males were substantially under-represented in all six states. The least educated (Did not graduate from high school) were under-represented and the highly educated (Graduated from college or technical school) were over-represented.

### 3. Identifying Factors Related To Key Survey Outcome Variables

Our current work relates to identifying socio-demographic factors associated with 13 key risk factor and health condition dichotomous outcome variables in the 2003 BRFSS. These include general health status, health insurance coverage, current smoking status, diabetes, and asthma. We used the same socio-demographic variables examined by Rao et al.: age group, gender, race/ethnicity, marital status, education, employment status, number of children in the household, and number of adults in the household.

Using the forward stepwise logistic regression procedure available in SAS Version 8.2, 13 weighted risk factor models were run to determine which socio-demographic variables were the best predictors of the risk factors. We considered independent predictor variables that entered at the first, second, or third step as the main predictors. Age entered all 13 models in the first, second, or third step. Education and race/ethnicity also entered most of the models. Marital status and gender entered four and three models, respectively.

Furthermore, we identified two-way interactions using weighted CHAID segmentation trees. We first collapsed some of the categories of the above five predictor variables: 1) age was collapsed into three categories (18-34, 35-54, and 55+), 2) education was collapsed into two categories (high school graduate or less, some college or more), and race/ethnicity was collapsed into three categories (non-Hispanic white and other races, non-Hispanic black, and Hispanic). Age by education was a key two-factor interaction in four of the CHAID models. Age by gender was a key two-factor interaction in 3 of the 13 CHAID models. Age by race/ethnicity was a key two-factor interaction in two of the CHAID models.

#### 4. Adding Variables To The BRFSS Weighting Methodology

The 2003 BRFSS weighting methodology involves calculating a base sampling weight (design weight) followed by poststratification to 14 age (7 categories)-by-gender control totals or 28 age-by-gender-by-race/ethnicity (non-Hispanic white versus all other race/ethnicity groups) totals to obtain the final weight. The control totals are obtained from Claritas, Inc. Our objective was to rake the 2003 BRFSS for each of the six states to CPS control totals constructed using the March 2002, 2003, and 2004 CPS. We combined three years of CPS data to add stability to the state-level control totals.

As expected, the Claritas population distribution for age by gender or age by gender by race/ethnicity in a state did not agree exactly with the CPS distribution for 2003-2004. Before obtaining control totals from the CPS, we first took the CPS March supplement person weight for each year and divided it by three. We then ratio-adjusted the CPS weight for the 14 age-by-gender or 28 age-by-gender-by-race/ethnicity categories, so that the CPS-weighted counts agreed with the Claritas counts. This step was necessary because we wanted to compare the impact of adding variables to the BRFSS weighting with the results from using the final BRFF weight. Once we had a new CPS weight, control totals were produced for race/ethnicity, education, marital status, age by education, and age by race/ethnicity. For each state, we collapsed the race/ethnicity variable to combine small categories

that constituted less than 5% of the BRFSS completed interviews in the state with an appropriate race/ethnicity category.

The CPS also has a variable indicating whether the household in which the adult lives has telephone service, so in each state we can estimate the number of adults living in nontelephone households at the time of the CPS interview. The 2003 BRFSS contains a variable indicating whether the respondent lives in a household that has experienced an interruption in telephone service of a week or longer. Using the BRFSS design weight, we estimated the percentage of adults in a state living in telephone households with an interruption in telephone service. Following the procedure described by Frankel et al. (2003), we then created a CPS control total margin for: 1) adults in telephone households without an interruption in telephone service, and 2) adults in telephone households with an interruption in telephone service and adults living in nontelephone households. The inclusion of the nontelephone margin in the raking is intended to compensate for noncoverage from the exclusion of adults living in nontelephone households.

For each of the 13 risk factor outcome variables, we used the BRFSS design weight and the BRFSS final weight to estimate the percent of adults with a risk factor in each of the six states. We then used a SAS raking macro (Battaglia et al. 2004) to create 10 new weights for the BRFSS in each of the six states. The details of the margins included in each raking are shown in Table 1. The logic to the ordering of the 10 rakings is as follows: 1) the first 5 raking do not include a nontelephone adjustment using the interruption margin described above, 2) most survey statisticians would give highest priority including a detailed race/ethnicity margin, even if a state has an age-by-gender by race/ethnicity margin limited to non-Hispanic white versus all other race/ethnic groups, 3) based on the logistic regression modeling results, education will next be entered as a margin, followed by marital status, and 4) based on the CHAID results, the age-by-education two-variable margin will next be entered and finally the age-by-race/ethnicity two-variable margin will be entered into the raking.

#### 5. Results For The Six States

For illustrative purposes, we show the results of the 10 rakings for two states – California and Texas. California uses age-by-gender-by-race/ethnicity poststratification, and only 2.8% of its adults reside in nontelephone households according to the CPS. The Texas BRFSS used age-by-gender poststratification and a higher proportion of its adults (5.7%) reside in nontelephone households based on the CPS. The race/ethnicity margin that we created using the 5% rule for Texas contains three

categories – non-Hispanic white, non-Hispanic black, and Hispanic plus non-Hispanic other races. For California, the race/ethnicity margin contains four categories – non-Hispanic white, non-Hispanic black, Hispanic, and non-Hispanic other races. We show results only for the question about general health status (see Figures 1 and 2), but the findings for the other risk factor variables are similar.

In California, the addition of the race/ethnicity margin has a small effect on the general health risk factor estimate. The raking that includes race/ethnicity and adds education increases the risk factor estimate. The addition of marital status, age by education, and age by race/ethnicity causes little further change in the estimate. Furthermore, the inclusion of the nontelephone margin in the raking has little impact. Compared to the risk factor estimates based on the final weight, the risk factor estimate from raking #10, which includes the nontelephone margin and the age-by-race margin, increases by 9.9%.

In Texas, the addition of the race/ethnicity margin has a larger effect on the general health risk factor estimate. The raking that includes race/ethnicity and adds education further raises the estimate. The addition of marital status, age by education, and age by race/ethnicity causes a small additional change in the estimate. Furthermore, the inclusion of the nontelephone margin in the raking noticeably raises the risk factor estimate. Compared to the risk factor estimate based on the final weight, the risk factor estimate from raking #10, which includes the nontelephone margin and the age-by-race margin, increases by 14.9%.

We developed estimates of the mean squared error (MSE) of the risk factor estimates (based on the design weight, the final weight, and raking weights #1 to #9) by treating the estimates from raking #10 as unbiased. Relative MSE estimates were calculated by dividing the square root of the MSE by the risk factor estimate from raking #10. Finally, we indexed the relative MSE estimates to the relative MSE estimates resulting from the BRFSS design weight. The indexed relative MSE results for the general health risk factor estimate for California and Texas are shown in Figures 3 and 4. By definition, the indexed relative MSE for the design weight estimates is 100%. Because the inclusion of more variables in the raking typically increases the variance, it is possible for the indexed relative MSE for estimates based on one of the other weights to exceed 100%. For California, the estimate based on the final weight and those for raking #1 (includes race/ethnicity) yield a reduction in the indexed relative MSE. However, a large additional reduction is seen with the addition of education to the raking. The inclusion of the nontelephone adjustment margin in the raking has very little impact on the indexed relative MSE

in California. We see a similar pattern in Texas except in terms of the indexed relative MSE for the final weight and the raking that includes race/ethnicity. Similar to California, we see that the addition of education to the raking causes a large drop in the indexed relative MSE. However, unlike California, the inclusion of the nontelephone adjustment margin has a noticeable impact on further reducing the indexed relative MSE. For general health status, the value of the indexed relative MSE is around 30% for the raking that includes the nontelephone margin and the age by education margin (raking #9). The inclusion of education, a socioeconomic status variable, is clearly important; however, the inclusion of the nontelephone adjustment margin in the raking can also be important for bias reduction.

## 6. Applying The Raking Method To All States

Based on what we learned in the six states, a new weight was developed for each of the 50 states and the District of Columbia. Following the approach of using the 2002-2004 March CPS, the CPS weight for the adults in each state was ratio-adjusted to the Claritas age-by-gender or age-by-gender-by-race/ethnicity distribution. Raking margins were then developed for race/ethnicity, education, marital status, age by gender, age by education, age by race/ethnicity, and for the nontelephone adjustment. For the race/ethnicity margin, a category-collapsing procedure was used to ensure that each category had at least 5.0% of the completed BRFSS interviews. For the age-by-race/ethnicity margin, the race/ethnicity categories developed for the one-variable race/ethnicity margin were used and age categories were collapsed to ensure that each contained at least 5.0% of the completed BRFSS interviews. The risk factor estimates based on the raking weight were then compared with the estimates based on the BRFSS poststratified weight. In Figure 4, the BRFSS general health risk factor estimates for the 50 states and DC are given on the horizontal axis. The vertical axis shows the difference. All of the differences are at least zero, indicating that the raking leads to risk factor estimates that are higher than the usual BRFSS estimates, generally by 1 to 3 percentage points. Similar results were found for the other risk factor estimates.

## 7. Conclusions

Data based on self-reports from a telephone survey can lead to an under-estimation of some risk factors in the population. For many states response rates have fallen below 50%, increasing the potential for nonresponse bias. People with no telephone service tend to be of lower socio-economic status, a characteristic associated with increased risk factors. Moreover, as the use of cellular

telephones increases, another layer of complexity is added in producing valid survey estimates. The methodology presented here will ensure a better weighting mechanism to overcome these limitations. By identifying socio-demographic variables associated with key risk factor variables, which are also related to unit nonresponse, and including these variables in the weighting methodology, we were able to substantially reduce nonresponse bias in the state risk factor estimates. The inclusion of a nontelephone adjustment margin can also lead to noncoverage bias reduction in some states. We found that many of the risk factor estimates increased noticeably when these variables were incorporated into the weighting using raking. Indeed, weighting through simple poststratification by age-sex or age-sex-race may be obsolete, and there is a need to further expand the list of variables to be accounted for in weighting. The methodology outlined here will better ensure that telephone survey results more closely match those produced by higher response rate area probability surveys conducted in homes.

### References

- Battaglia, M.P., Izrael, D., Hoaglin, D.C., and Frankel, M.R. 2004. Tips and tricks for raking survey data. *2004 Proceedings of the Annual Meeting of the American Statistical Association* [CD-ROM], Alexandria, VA: American Statistical Association.
- Battaglia, M.P., Khare, M., Frankel, M.R., Murray, M.C., Buckley, P., and Peritz, S. 2006. Response rates: How have they changed and where are they headed? Monograph paper presented at the Telephone Survey Methodology II Conference, Miami, FL, January 2006.
- Curtin, R., Presser, S., and Singer, E. 2005. Changes in Telephone Survey Nonresponse Over the Past Quarter Century. *Public Opinion Quarterly*, Volume 69: 87-98.
- Frankel, M.R., Srinath, K.P., Hoaglin, D.C., Battaglia, M.P., Smith, P.J., Wright, R.A., and Khare, M. 2003. Adjustments for non-telephone bias in random-digit-dialling surveys. *Statistics in Medicine*, Volume 22, pp. 1611-1626.
- Rao, S.R., Link, M.W., Battaglia, M.P., Frankel, M.R., Giambo, P., and Mokdad, A.H. 2005. Assessing representativeness in RDD surveys: Coverage and nonresponse in the Behavioral Risk Factor Surveillance System. *2005 Proceedings of the Annual Meeting of the American Statistical Association* [CD-ROM], Alexandria, VA: American Statistical Association.

Table 1: Margins Included in the 10 BRFSS Rakings

Without interruption in telephone service margin:	
1. Age by gender or age by gender by race/ethnicity	And race/ethnicity
2. Age by gender or age by gender by race/ethnicity and race/ethnicity	And education
3. Age by gender or age by gender by race/ethnicity, race/ethnicity, education	And marital status
4. Age by gender or age by gender by race/ethnicity, race/ethnicity and marital status	And age by education
5. Age by gender or age by gender by race/ethnicity, race/ethnicity and age by education	And age by race/ethnicity
With interruption in telephone service margin:	
6. Age by gender or age by gender by race/ethnicity	And race/ethnicity and interruption in telephone service
7. Age by gender or age by gender by race/ethnicity and race/ethnicity	And education and interruption in telephone service
8. Age by gender or age by gender by race/ethnicity, race/ethnicity, education	And marital status and interruption in telephone service
9. Age by gender or age by gender by race/ethnicity, race/ethnicity and marital status	And age by education and interruption in telephone service
10. Age by gender or age by gender by race/ethnicity, race/ethnicity and age by education	And age by race/ethnicity and interruption in telephone service

Figure 1: Graph of California General Health Risk Factor Estimates for BRFSS Poststratified Weight and 10 Raking Weights

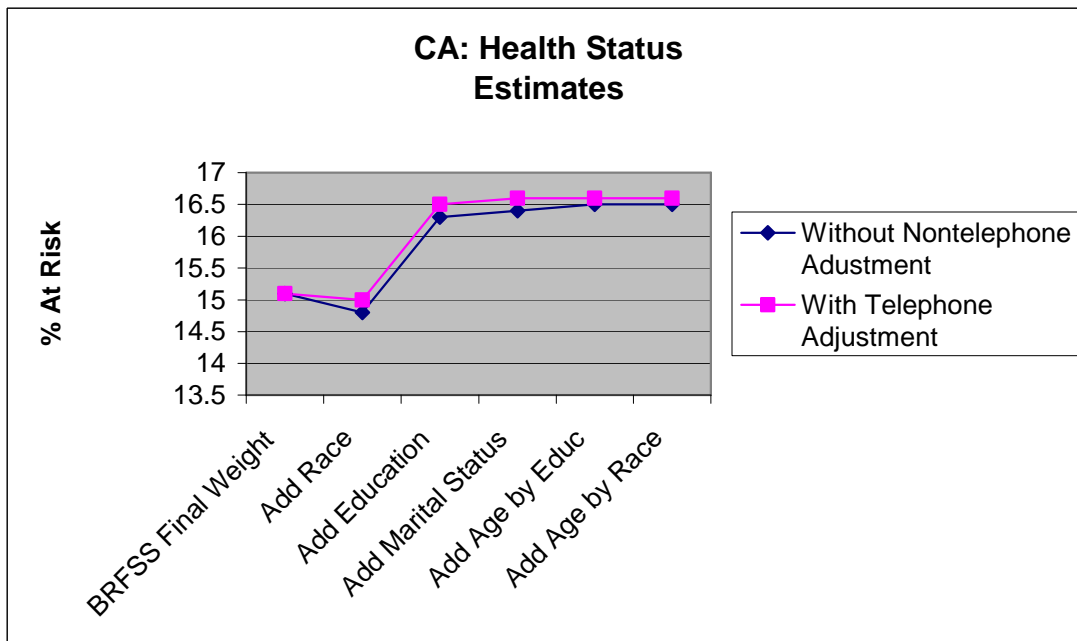


Figure 2: Graph of Texas General Health Risk Factor Estimates for BRFSS Poststratified Weight and 10 Raking Weights

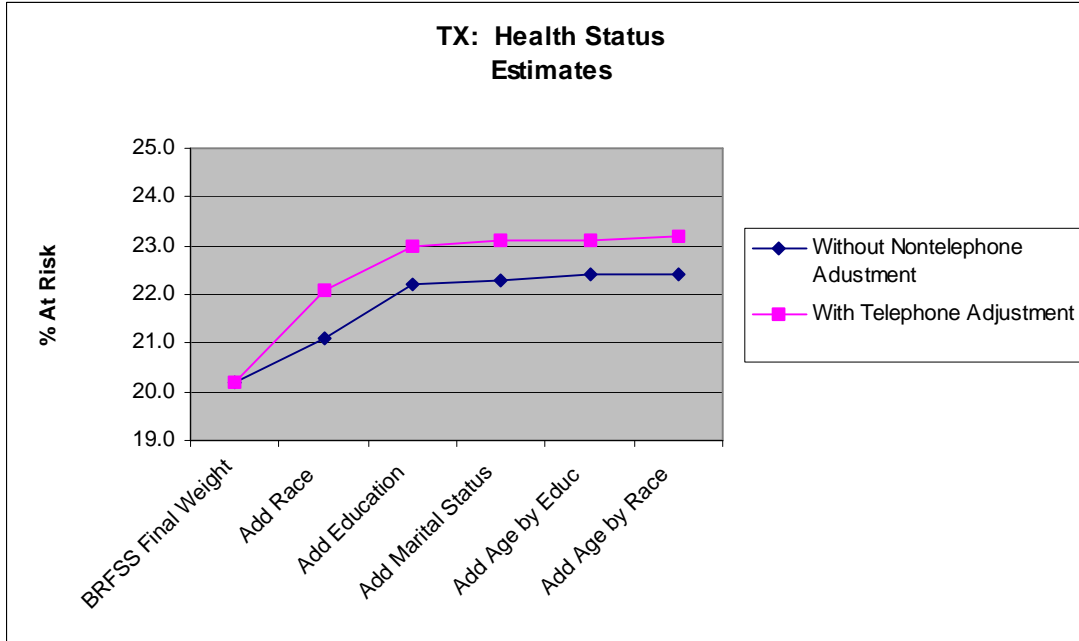


Figure 3: Graph of Indexed Relative Mean Squared Error for California General Health Risk Factor Estimates

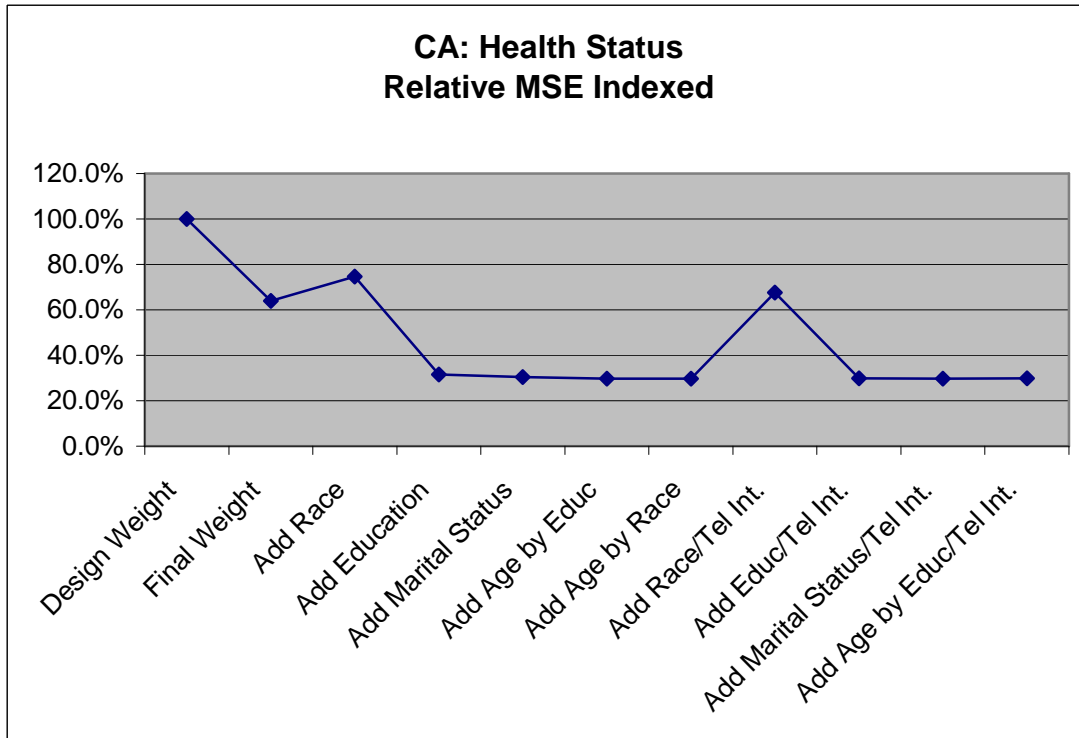


Figure 4: Graph of Indexed Relative Mean Squared Error for Texas General Health Risk Factor Estimates

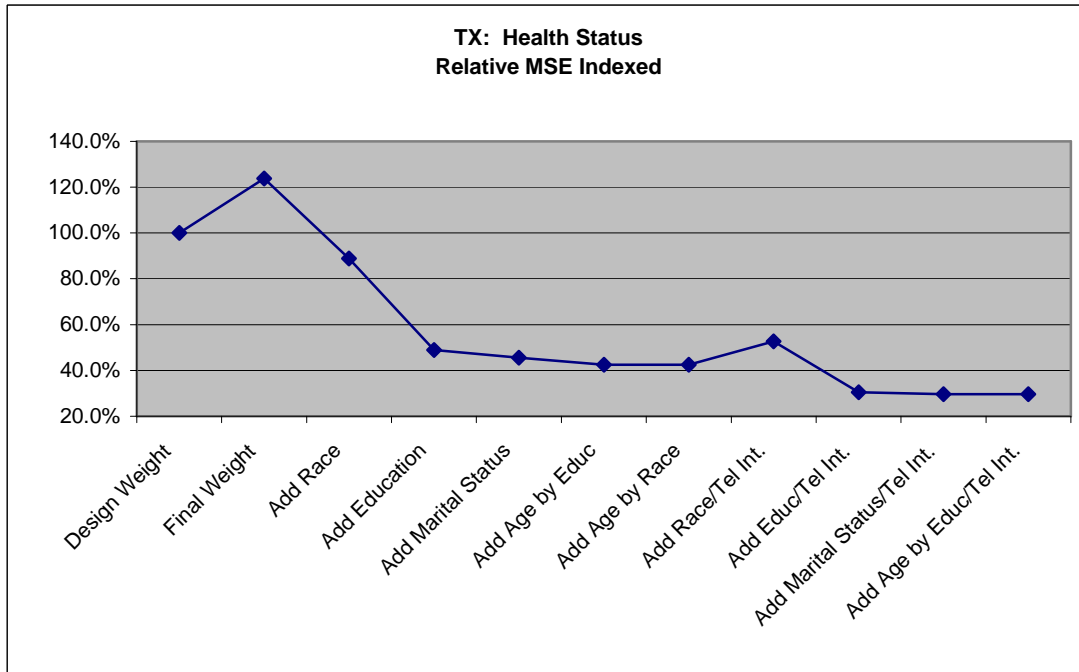


Figure 5: Difference in General Health Risk Factor Estimates (Raking – Poststratification) Plotted Against BRFSS General Health Risk Factor Estimate for 50 States and DC

