Appendix G

Mathematical Proof of Why the Potential for Bias Resulting from not Refreshing our Sample to Include “New Births” in 2004 is Small

Assume that we are estimating a population proportion of some characteristic of interest (yes/no type) relating to school districts. We can think of the population of school districts as being divided into two strata. The first stratum would consist of school districts that existed in 1997 and the second stratum consists of school districts that existed in 2004 but not in 1997 ("new births"). Let the number of school districts in the population be N. Let the number of school districts in the first stratum be N_r and the number of school districts in the second stratum be N_m. We have:

$$N = N_r + N_m$$

The overall population proportion of interest can be written as a weighted average of the proportion among existing districts and the population proportion among new births. Let P be the overall proportion, P_r the proportion among existing districts and P_m the proportion among new births. The overall proportion can be written as:

$$P = \frac{N_r P_r + N_m P_m}{N}$$

The sample proportion we have in 2004 can only be computed from the responding school districts that existed in 1997. Let this proportion be p_r.

The bias in the sample proportion p_r because of not having any data from the new school districts is:

$$B(p_r) = E(p_r) - P$$

The bias in the estimate is the difference between the expected value of the estimate and population proportion. The expected value is the average of sample proportions of all possible samples that we can draw from the population of respondents. We have

$$E(p_r) = P_r$$

Therefore, the bias in the estimate is

$$B(p_r) = P_r - P$$

That is, the bias is the difference between the proportion among the respondents minus the overall proportion. This can be written as:

$$B(p_r) = P_r - \frac{N_r P_r + N_m P_m}{N}$$

Alternatively, this can be written as:

$$B(p_r) = \frac{NP_r - N_r P_r - N_m P_m}{N}$$

Since we have $N = N_r + N_m$, we can write $B(p_r)$ as

$$B(p_r) = \frac{N_m}{N}(P_r - P_m)$$

The bias in the estimate due to not including the new births is small if either (1) $\frac{N_m}{N}$ which is the proportion of new births is small or (2) the difference between the proportion among existing districts and the proportion among new births is small.

In this case $\frac{N_m}{N}$ is small (2%). Therefore, the bias is not likely to be large as this difference gets multiplied by a small number.