APPENDIX G
ESTIMATING SAMPLING ERRORS

The final adjusted weights are used in calculating parameter estimates and their sample variances. RTI uses SUDAAN for these calculations. Variance estimation for the statistics computed in the SUDAAN series of procedures for survey data analysis is based on a first-order Taylor series approximation of the deviations of estimates from their expected values. This approximation for large samples is well-known (see Kendall and Stuart, 1961, p. 231). Woodruff (1971) presented applications of this technique to sample surveys. This method yields one of the best known numerical approximations currently available in the statistical literature for ratio estimates. The general approach taken to compute variances is to first form the Taylor series linearization for a particular statistic. These linearized values are referred to as Z_i for the ith sample unit throughout this appendix. Once the linearized values are formed, they are substituted into the formula for computing the variance of a total estimate that is appropriate for the design.

Estimating the total number of individuals who belong to an arbitrarily defined domain or subpopulation provides a convenient example. Denote the total in question by $\hat{N}_d$, where d denotes the domain. Establish a domain indicator

$$I_{hijk} = \begin{cases} 1 \text{ if the } k^{th} \text{ person is in the domain} \\ 0 \text{ if the } k^{th} \text{ person is not in the domain} \end{cases}$$

where

h is the stratum, h = 1, ..., H
i is the ith cluster, in stratum h, i=1, ..., n_h
j is the jth household in the cluster i in stratum h, j=1, ..., n_i
k is the kth person in the household; in cluster i in stratum h, k=1, ..., n_j

and w_{hijk} is the population weight for person k in household j in cluster i in stratum h.

Then, $z_{hijk} = I_{hijk} \cdot w_{hijk}$

and the estimate of the domain total is

$$\hat{N}_d = \sum_h \sum_i \sum_j \sum_k z_{hijk}$$

and the variance of this estimate is

$$Var(z) = \sum_h n_h s_h^2$$

where

$$s_h^2 = \frac{\sum_i \left(z_{hi} - \bar{z}_h\right)^2}{n_h - 1}$$, the stratum-level sum of squares,

with

$$z_{hi} = \sum_j \sum_k z_{hijk}$$, the cluster-level sum,

and

$$\bar{z}_h = \frac{\sum_i z_{hi}}{n_h}$$, the stratum-level mean.

Other methods of obtaining the variance estimates could be used instead of the first order Taylor series linearizations. Examples include such pseudorandomization techniques as balanced repeated replications (BRR), jackknifing and boot-strapping. The Taylor series linearization is preferred by many because of its computational efficiency (generally less demanding of computer time).

The most commonly used statistical packages, such as SAS, BMDP, and SPSS, do <u>not</u> calculate standard errors of survey estimates accounting for complex sample designs. There are, however, several commercially available packages that can correctly calculate the standard errors for designs such as the one used in NPTS, among them are:

- Clusters (World Health Organization
- Osiris (University of Michigan)
- SUDAAN (RTI)
- Super Carp (Iowa State University)
- Wesvar Procedures (Westat)

Of these, all use Taylor series linearization except Wesvar, which uses BRR.