Survey Methodology, December 1996
Vol. 22, No. 2, pp. 157-165
Statistics Canada

Stability Measures for Variance Component Estimators Under a Stratified Multistage Design

J.L. ELTINGE and D.S. JANG[1]

ABSTRACT

In work with sample surveys, we often use estimators of the variance components associated with sampling within and between primary sample units. For these applications, it can be important to have some indication of whether the variance component estimators are stable, *i.e.*, have relatively low variance. This paper discusses several data-based measures of the stability of design-based variance component estimators and related quantities. The development emphasizes methods that can be applied to surveys with moderate or large numbers of strata and small numbers of primary sample units per stratum. We direct principal attention toward the design variance of a within-PSU variance estimator, and two related degrees-of-freedom terms. A simulation-based method allows one to assess whether an observed stability measure is consistent with standard assumptions regarding variance estimator stability. We also develop two sets of stability measures for design-based estimators of between-PSU variance components and the ratio of the overall variance to the within-PSU variance. The proposed methods are applied to interview and examination data from the U.S. Third National Health and Nutrition Examination Survey (NHANES III). These results indicate that the true stability properties may vary substantially across variables. In addition, for some variables, within-PSU variance estimators appear to be considerably less stable than one would anticipate from a simple count of secondary units within each stratum.

KEY WORDS: Between-PSU variance; Complex sample design; Degrees of freedom; Diagnostic; Design-based analysis; Satterthwaite approximation; Stratum collapse; U.S. Third National Health and Nutrition Examination Survey (NHANES III); Within-PSU variance.

1. INTRODUCTION

In work with sample surveys, it is often desirable to have good estimates of the variance components attributable to sampling within and between primary sample units (PSUs). For example, the magnitude of an estimated within-PSU variance, relative to a between-PSU variance, may influence decisions regarding sample allocation and related design issues (*e.g.*, Hansen *et al.* 1953, Chapter 7). Similar relative-magnitude properties affect the bias of certain variance estimators derived under simplifying assumptions regarding the sample design (*e.g.*, Korn and Graubard 1995, p. 278-279, 287; and Wolter 1985, p. 44-46). Also, some survey analysts have a general interest in identification of surveys and variables for which the between-PSU component of variance is substantially greater than zero. See, *e.g.*, Herzog and Scheuren (1976, p. 398) and Wolter (1985, p. 47) for related comments. In addition, Jang and Eltinge (1996) give an example for which there is some interest in the within-PSU variances by themselves.

In some application work, estimates of within-PSU variances and related quantities are reported with the apparent assumption that the estimates are stable, *i.e.*, have relatively low variances. This paper shows that it can be important to carry out data-based checks of this assumption of stability, and that some relatively simple checking methods follow from standard design-based ideas. We emphasize methods that can be applied to designs with a moderate or large number of strata and a small number of PSUs selected per stratum.

Subsection 2.1 reviews the relevant estimators of within-PSU variances and overall stratum-level variances. Subsection 2.2 identifies two distinct components of the variance of the within-PSU variance estimator. Subsection 2.3 presents simple design-based estimators of the variances of two within-PSU variance estimators. Section 3 develops two related degrees-of-freedom measures.

Section 4 examines the extent to which related design-based methods can be used to assess the stability of quantities that depend both on the within-PSU variance estimator and on the overall stratum-level variance estimator. Principal attention is directed toward an estimator of the between-PSU variance and an estimator of the ratio of the overall stratum-level variance divided by the within-PSU variance. Section 4.2 discusses one set of methods based on the stability measures from Section 2 and some moderately restrictive moment assumptions. Section 4.3 outlines a second set of methods based on stratum collapse.

Section 5 applies the main ideas of Sections 2 through 4 to variance estimates computed for the U.S. Third National Health and Nutrition Examination Survey. Section 5 also uses a simple simulation-based method to assess the consistency of the observed measures with standard assumptions regarding variance estimator stability. The Section 5 results suggest that the true stability of within-PSU variance estimators can be substantially less than anticipated from a simple count of the number of secondary units contributing to each PSU. In addition, the results indicate that the stability properties of

[1] J.L. Eltinge and D.S. Jang, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A.

within-PSU variance estimators and related quantities can vary substantially across different variables collected in the same survey. Section 6 gives additional comments on the methods and empirical results presented here.

2. WITHIN-PSU AND OVERALL STRATUM-LEVEL VARIANCE ESTIMATORS

2.1 General Notation

In principle, we could use either design-based or model-based methods to examine within-PSU and between-PSU variance components. The present work will take a design-based approach. This is consistent with some related previous literature, *e.g.*, Wolter (1985, p. 40-41, 47). The design-based approach will be especially useful in highlighting some strengths and limitations of the proposed stability-assessment methods. For example, in Section 2.3 this approach will give us some indication of specific design features that may affect variance estimator stability. Also, in Section 4 the design-based approach will help to clarify the extent to which certain moment restrictions are needed to justify one set of stability measures.

Following the notation and ideas in Wolter (1985, p. 43-47), consider a stratified multistage sample design with L strata and with N_h primary sampling units (PSUs) contained in stratum $h = 1, 2, ..., L$. We select n_h PSUs with replacement and with per-draw selection probabilities p_{hi}. Within selected PSU (h,i), we select n_{hi} secondary sample units (SSUs) with replacement and with per-draw selection probabilities p_{hij}. Further subsampling is carried out within a selected SSU to obtain n_{hij} individual elements for interview or examination. The stability-assessment methods developed here are intended primarily for designs with moderate or large L, relatively small n_h (*e.g.*, $n_h = 2$), and relatively large n_{hi}. Designs with these characteristics are often used in large household interview surveys, *e.g.*, the health survey discussed in Section 4.

We will focus on estimation of a population total $Y = \sum_{h=1}^{L} Y_h$, where $Y_h = \sum_{i=1}^{N_h} Y_{hi}$, $Y_{hi} = \sum_{j=1}^{N_{hi}} \sum_{k=1}^{N_{hij}} Y_{hijk}$, Y_{hijk} is a survey item for element k in SSU j in PSU i in stratum h, N_{hi} is the number of SSUs in PSU (h,i), and N_{hij} is the number of elements in SSU (h, i, j). Extensions to nonlinear functions of population totals are straightforward and will be considered in the applications in Section 5. A standard design-based estimator of Y is $\hat{Y} = \sum_{h=1}^{L} \hat{Y}_h$ where

$$\hat{Y}_h = \sum_{i=1}^{n_h} \sum_{j=1}^{n_{hi}} \sum_{k=1}^{n_{hij}} w_{hijk} y_{hijk}, \qquad (2.1)$$

w_{hijk} is the customary weight derived from selection probabilities and sample sizes to ensure unbiased estimation of each Y_h, and the lower-case terms y_{hijk} denote sample observations. In subsequent work, it will be useful to rewrite expression (2.1) as

$$\hat{Y}_h = n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} \hat{Y}_{hi},$$

where $\hat{Y}_{hi} = n_{hi}^{-1} \sum_{j=1}^{n_{hi}} z_{hij}$ and $z_{hij} = n_h n_{hi} p_{hi} \sum_{k=1}^{n_{hij}} w_{hijk} y_{hijk}$.

2.2 Within- and Between-PSU Variances

Throughout this discussion, expectations and variances will be defined with respect to the sample design. Under the conditions stated above, the variance of $\hat{Y}$ is $V(\hat{Y}) = \sum_{h=1}^{L} V_h$, where $V_h = V_{Bh} + V_{Wh}$, $V_{Bh} = V(n_h^{-1} \sum_{i=1}^{n_h} p_{hi}^{-1} Y_{hi})$, $V_{Wh} = n_h^{-1} \sum_{i=1}^{N_h} p_{hi}^{-1} \sigma_{2hi}^2$, and $\sigma_{2hi}^2 = V(\hat{Y}_{hi} - Y_{hi} \mid h,i)$; see, *e.g.*, Wolter (1985, p. 42). Note especially that Y_{hi} is the true population total for selected PSU (h, i), and that σ_{2hi}^2 reflects the variability in $\hat{Y}_{hi} - Y_{hi}$ attributable to subsampling at the SSU and finer levels.

A customary unbiased estimator of the overall stratum-level variance V_h is

$$\hat{V}(\hat{Y}_h) = n_h^{-1} (n_h - 1)^{-1} \sum_{i=1}^{n_h} (p_{hi}^{-1} \hat{Y}_{hi} - \hat{Y}_h)^2,$$

and the corresponding estimator of $V(\hat{Y}) = \sum_{h=1}^{L} V(\hat{Y}_h)$ is $\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \hat{V}(\hat{Y}_h)$.

Now consider estimation of the within-PSU variance V_{Wh}. Since $\hat{Y}_{hi}$ is a sample mean of the independent and identically distributed terms z_{hij}, standard arguments show that for a given PSU (h,i), an unbiased estimator of σ_{2hi}^2 is $\hat{\sigma}_{2hi}^2 = n_{hi}^{-1} (n_{hi} - 1)^{-1} \sum_{j=1}^{n_{hi}} (z_{hij} - \hat{Y}_{hi})^2$. Thus, an unbiased estimator of V_{Wh} is

$$\hat{V}_{Wh} = \sum_{i=1}^{n_h} n_h^{-2} p_{hi}^{-2} \hat{\sigma}_{2hi}^2 = \sum_{i=1}^{n_h} n_{hi}^{-1} (n_{hi} - 1)^{-1} \sum_{j=1}^{n_{hi}} (x_{hij} - \overline{x}_{hi})^2,$$

where $x_{hij} = n_{hi} \sum_{k=1}^{n_{hij}} w_{hijk} y_{hijk}$ and $\overline{x}_{hi} = n_{hi}^{-1} \sum_{j=1}^{n_{hi}} x_{hij}$. Note that the latter expression for $\hat{V}_{Wh}$ uses only sample sizes, the observations y_{hijk} and the customary weights w_{hijk}.

2.3 The Variance of $\hat{V}_{Wh}$

A direct modification of standard conditional-moment arguments shows that the variance of $\hat{V}_{Wh}$ is $\gamma_{Bh} + \gamma_{Wh}$, where

$$\gamma_{Bh} = V(n_h^{-2} \sum_{i=1}^{n_h} p_{hi}^{-2} \sigma_{2hi}^2)$$

and

$$\gamma_{Wh} = n_h^{-3} \sum_{i=1}^{N_h} p_{hi}^{-3} V(\hat{\sigma}_{2hi}^2 \mid h,i).$$

Thus, the variance of $\hat{V}_{Wh}$ itself depends on a sum of between- and within-PSU variances, and the relative magnitudes of γ_{Bh} and γ_{Wh} depend on trade-offs among σ_{2hi}^2, p_{hi} and n_{hi}. For example, under regularity conditions, the terms $V(\hat{\sigma}_{2hi}^2 \mid h,i)$ are approximately inversely proportional to n_{hi}. Thus, if the n_{hi} are uniformly large within stratum h, then γ_{Wh}

may be relatively small. Also, if the terms $p_{hi}^{-2}\sigma_{2hi}^2$ are approximately constant within a given stratum, then γ_{Bh} may be relatively small. Conversely, marked heterogeneity of $p_{hi}^{-2}\sigma_{2hi}^2$ may inflate γ_{Bh} and thus inflate $V(\hat{V}_{Wh})$ as well.

In addition, note that under the stated design conditions, $\hat{V}_{Wh}$ is the sample mean of the independent and identically distributed terms $n_h^{-1}p_{hi}^{-2}\hat{\sigma}_{2hi}^2$. Thus, an unbiased estimator of the variance of $\hat{V}_{Wh}$ is

$$\tilde{V}(\hat{V}_{Wh}) = n_h^{-1}(n_h - 1)^{-1}\sum_{i=1}^{n_h}(n_h^{-1}p_{hi}^{-2}\hat{\sigma}_{2hi}^2 - \hat{V}_{Wh})^2. \qquad (2.2)$$

Some applications focus on the full-population level, rather than on individual strata, and so the "within-PSU" contribution of interest is the sum of the within-PSU variances, $V_W = \sum_{h=1}^{L} V_{Wh}$. Under the conditions given above, an unbiased estimator of V_W is $\hat{V}_W = \sum_{h=1}^{L}\hat{V}_{Wh}$. Also, since our sampling and subsampling are independent across strata, we have $V(\hat{V}_W) = \sum_{h=1}^{L}(\gamma_{Bh} + \gamma_{Wh})$, and an unbiased estimator of $V(\hat{V}_W)$ is

$$\tilde{V}(\hat{V}_W) = \sum_{h=1}^{L}\tilde{V}(\hat{V}_{Wh}).$$

Finally, note that the preceding development used the assumption of sampling with replacement at both the primary- and secondary-unit levels. Two applications of result (2.4.16) in Wolter (1985, p. 46) show that under mild conditions that hold for many, but not all, without-replacement designs, $\hat{V}_{Wh}$ will be unbiased or conservative for the true within-PSU variance; and $\tilde{V}(\hat{V}_{Wh})$ will be unbiased or conservative for the true variance of $\hat{V}_W$. A formal technical statement and proof of this result is available from the authors.

2.4 Balanced Interpretation of Stability Measures

The remainder of this paper uses $\tilde{V}(\hat{V}_{Wh})$ and related quantities to assess the stability of variance-component estimators. In working with these results, it is useful to remember that data-based measures of variance estimator stability are justifiably viewed with some caution, because they are functions of fourth sample moments, and thus are themselves subject to a considerable amount of sampling variability. See, *e.g.*, Fuller (1984, p. 111). This caution carries over to the proposed estimator $\tilde{V}(\hat{V}_{Wh})$ and to the related statistics discussed in Sections 3 and 4 below.

However, one should not overstate this caution to the point of making no attempt at data-based assessment of variance estimator stability. The estimator $\tilde{V}(\hat{V}_{Wh})$, and the related measures in Sections 3 and 4, are relatively simple to compute, and provide diagnostics that can help to identify variables for which:

(a) the instability of $\hat{V}_{Wh}$ is especially problematic; or

(b) the instability of $\hat{V}_{Wh}$ has a substantial effect on the precision of estimators of the relative magnitudes of between-PSU and within-PSU variances.

Consequently, interpretation of specific values of $\tilde{V}(\hat{V}_{Wh})$ and related stability measures should reflect a balance between the abovementioned general caution and a recognition of their potential diagnostic value.

3. TWO STABILITY MEASURES FOR WITHIN-PSU VARIANCE ESTIMATORS

3.1 Degrees-of-Freedom Diagnostics for Variance Estimator Stability

Some analysts prefer to express variance estimator stability through "degrees of freedom" measures related to the Satterthwaite (1941, 1946) approximation. To introduce this idea, consider a general variance estimator $\hat{V}$, and note that $\{E(\hat{V})\}^{-1}d\hat{V}$ has the same first and second moments as a chi-square random variable on d degrees of freedom, where d is the solution to the equation,

$$2\{E(\hat{V})\}^2 - V(\hat{V})d = 0.$$

If the distribution of $\{E(\hat{V})\}^{-1}d\hat{V}$ is indeed well approximated by a chi-square distribution, then d may be viewed fairly literally as a "degrees of freedom" term. Otherwise, d can be viewed as twice the inverse of the squared coefficient of variation of $\hat{V}$. In either case, d has a certain appeal because it is scale-free, and can be tied fairly directly to notions of "effective sample size" in the evaluation of variance estimator performance. Subsection 3.3 gives related comments for two special cases.

Given an unbiased estimator $\tilde{V}(\hat{V})$ of the variance of $\hat{V}$, one may compute a "degrees of freedom" estimator $\hat{d}$ as the solution to the unbiased estimating equation

$$2\{\hat{V}^2 - \tilde{V}(\hat{V})\} - \tilde{V}(\hat{V})d = 0, \qquad (3.1)$$

i.e., $\hat{d} = \{\tilde{V}(\hat{V})\}^{-1}2\hat{V}^2 - 2$. Under mild regularity conditions, $d^{-1}\hat{d}$ converges in probability to one, provided $\{V(\hat{V})\}^{-1}\tilde{V}(\hat{V})$ and $\{E(\hat{V})\}^{-1}\hat{V}$ both converge in probability to one.

3.2 Degrees-of-Freedom Diagnostics for Pooled and Stratum-Level Estimators of Within-PSU Variances

We can apply these general degrees-of-freedom ideas to the within-PSU variance estimators $\hat{V}_{Wh}$ and $\hat{V}_W$ developed in Section 2. First consider the case in which there is intrinsic interest in the stability of individual stratum-level estimators $\hat{V}_{Wh}$. The associated "degrees of freedom" measure is $d_{Wh} = \{V(\hat{V}_{Wh})\}^{-1}2V_{Wh}^2$. For designs with large n_h, one may use (3.1) to compute estimators $\hat{d}_{Wh} = \{\tilde{V}(\hat{V}_{Wh})\}^{-1}2\hat{V}_{Wh}^2 - 2$ separately for each stratum. For designs with small n_h (*e.g.*, $n_h = 2$ for each stratum), the estimator $\hat{d}_{Wh}$ itself may be very unstable.

Consequently, it also is useful to consider the alternative combined estimator

$$\hat{d}_{W0} = \left\{\sum_{h=1}^{L} \tilde{V}(\hat{V}_{Wh})\right\}^{-1} 2\sum_{h=1}^{L} \hat{V}_{Wh}^{2} - 2,$$

under the assumption that all d_{Wh} equal a common value d_{W0}.

Now consider the pooled within-PSU variance estimator $\hat{V}_W$ developed in Section 2.3. The resulting "degrees of freedom" measure is $d_{WF} = \{\sum_{h=1}^{L} V(\hat{V}_{Wh})\}^{-1} 2V_W^2$, and expression (3.1) suggests the estimator

$$\hat{d}_{WF} = \left\{\sum_{h=1}^{L} \tilde{V}(\hat{V}_{Wh})\right\}^{-1} 2\hat{V}_W^2 - 2.$$

3.3 Comparison of d_{W0} and d_{WF} to Direct SSU Counts

To interpret $\hat{d}_{W0}$ and $\hat{d}_{WF}$ as stability measures, consider the following idealized setting. Assume that for all h, the PSU counts n_h are equal to a common value n_1, say; and that for all h and i, the SSU counts n_{hi} are equal to a common value n_{11}. In addition, assume that the terms $p_{hi}^{-2}\sigma_{2hi}^{2}$ are constant within each stratum; and that, conditional on (h, i), each $\sigma_{2hi}^{-2}(n_{11} - 1)\,\hat{\sigma}_{2hi}^{2}$ is distributed as a chi-square random variable on $n_{11} - 1$ degrees of freedom. Then routine arguments show that $d_{W0} = n_1(n_{11} - 1)$. If the preceding assumptions are satisfied approximately, and if the product $n_1(n_{11} - 1)$ is large (greater than 40, say), then a data user may be inclined to view $\hat{V}_{Wh}$ as relatively stable, or equivalently, to view the errors $\hat{V}_{Wh} - V_{Wh}$ as negligible. This appears to be the reasoning used implicitly when estimates $\hat{V}_{Wh}$ are treated as known values in design or analysis work. However, the application in Section 5 will give some examples for which this reasoning is problematic, so that evaluation of the estimates $\hat{d}_{W0}$ is important.

Also, under the idealized conditions described above, and under the additional assumption that the V_{Wh} are all equal, we have $d_{WF} = Ln_1(n_{11} - 1)$.

4. COMPARISON OF WITHIN-PSU AND OVERALL STRATUM-LEVEL VARIANCES

4.1 Estimators of Between-PSU Variances and Related Variance Ratios

Section 1 cited some applications that hinge on the magnitude of V_{Wh} relative to V_h. The specifics of the relative-magnitude comparisons vary with the individual application, but interest generally focuses on differences or ratios. For example, recall that $V_{Bh} = V_h - V_{Wh}$ and define the overall between-PSU variance term $V_B = \sum_{h=1}^{L} V_{Bh}$. In addition, note that unbiased estimators of V_{Bh} and V_B are $\hat{V}_{Bh} = \hat{V}_h - \hat{V}_{Wh}$ and $\hat{V}_B = \sum_{h=1}^{L} \hat{V}_{Bh}$ respectively.

Similarly, define the ratio $R_{WV} = V_W^{-1} V(\hat{Y})$, the magnitude of the overall variance $V(\hat{Y})$ relative to the within-PSU contribution V_W. A direct estimator of R_{WV} is $\hat{R}_{WV} = \hat{V}_W^{-1}\hat{V}(\hat{Y})$. Note that if $V_{Wh}^{-1}V_h = R_{WV}$ for all h, then $\hat{R}_{WV}$ could also be viewed as a pooled estimator of this common *stratum*-level ratio.

For both $\hat{V}_B$ and $\hat{R}_{WV}$, stability assessment involves the variance of $\hat{V}_h$ and the covariance of $\hat{V}_{Wh}$ with $\hat{V}_h$. Estimation of the these moments can be somewhat problematic for surveys that select small numbers of PSUs from each stratum. We consider two approaches to resolving this problem. Section 4.2 uses moderate restrictions on the moment structure of $(\hat{V}_{Wh}, \hat{V}_h)$ to develop estimators $V(\hat{V}_h)$ and related quantities. Section 4.3 uses stratum collapse to develop alternative stability measures.

4.2 Stability Measures Based on $\tilde{V}(\hat{V}_{Wh})$ and Moment Conditions

4.2.1 Moment Conditions

Under moderate moment restrictions, we can estimate the variance of $\hat{V}_h$ directly from $\hat{V}_h$ itself. Specifically, assume that the variance of $\hat{V}_h$ equals $(n_h - 1)^{-1}2V_h^2$; this would hold, *e.g.*, under the standard assumption that $V_h^{-1}(n_h - 1)\hat{V}_h$ is distributed as a chi-square random variable on $n_h - 1$ degrees of freedom. As in Sections 2 and 3, we continue to assume that $\hat{V}_h$ is unbiased for V_h. Then routine moment arguments show that $(n_h + 1)^{-1}2\hat{V}_h^2$ is an unbiased estimator of the variance of $\hat{V}_h$.

In the remainder of Section 4.2, we will also assume that $\mathrm{Cov}(\hat{V}_{Wh}, \hat{V}_h) = 0$. Routine conditional-moment arguments show that this will hold if the terms $p_{hi}^{-2}\sigma_{2hi}^{2}$ are equal within a given stratum; and if, conditional on (h,i,j), the SSU-level estimates x_{hij} are normally distributed, so that $\hat{\sigma}_{2hi}^{2}$ is conditionally independent of $\hat{Y}_{hi}$.

4.2.2 Stability Measures

Under the conditions stated in Section 4.2.1, unbiased estimators of $V(\hat{V}_{Bh})$ and $V(\hat{V}_B)$ are

$$\tilde{V}(\hat{V}_{Bh}) = (n_h + 1)^{-1}2\hat{V}_h^2 + \tilde{V}(\hat{V}_{Wh}) \qquad (4.1)$$

and $\tilde{V}(\hat{V}_B) = \sum_{h=1}^{L} \tilde{V}(\hat{V}_{Bh})$, where $\tilde{V}(\hat{V}_{Wh})$ is defined in expression (2.2). Also, under the same conditions routine ratio-estimation arguments lead to the variance estimator

$$\tilde{V}(\hat{R}_{WV}) = \hat{V}_W^{-2}\sum_{h=1}^{L}\left\{(n_h + 1)^{-1}2\hat{V}_h^2 + \hat{R}_{WV}^2\,\tilde{V}(\hat{V}_{Wh})\right\}. \qquad (4.2)$$

4.3 Alternative Stability Measures Based on Stratum Collapse

The assumptions of Section 4.2.1 may be problematic in some applications. For example, for some survey designs and variables, the SSU-level estimators x_{hij} may have markedly nonnormal distributions, so the assumption $\mathrm{Cov}(\hat{V}_{Wh}, \hat{V}_h) = 0$ may not hold. For these cases, one may consider the use of stratum collapse to produce alternative estimators of $V(\hat{V}_B)$ and $V(\hat{R}_{WV})$.

Specifically, partition the set of L strata into G prespecified groups, with L_g strata contained in group S_g, $g = 1, ..., G$. With this new notation, note that

$$(\hat{V}(\hat{Y}), \hat{V}_W, \hat{V}_B) = \sum_{g=1}^{G} \sum_{h \in S_g} (\hat{V}_h, \hat{V}_{Wh}, \hat{V}_{Bh}).$$

Standard stratum-collapse methods (*e.g.*, Wolter 1985, Section 2.5) then lead to the alternative variance estimator,

$$V_{cs}^{*}(\hat{V}_B) = \sum_{g=1}^{G} (L_g - 1)^{-1} L_g \sum_{h \in S_g} D_{gh}^2,$$

where $D_{gh} = \hat{V}_{Bh} - L_g^{-1} \sum_{j \in S_g} \hat{V}_{Bj}$. Similarly, a collapsed-stratum variance estimator for $\hat{R}_{WV}$ is,

$$V_{cs}^{*}(\hat{R}_{WV}) = (\hat{V}_W)^{-2} \sum_{g=1}^{G} (L_g - 1)^{-1} L_g \sum_{h \in S_g} C_{gh}^2,$$

where $C_{gh} = (\hat{V}_h - \hat{R}_{WV} \hat{V}_{Wh}) - L_g^{-1} \sum_{j \in S_g} (\hat{V}_j - \hat{R}_{WV} \hat{V}_{Wj})$.

In general, collapsed-stratum variance estimators require some care in interpretation; see, *e.g.*, Rust and Kalton (1985), Wolter (1985, Section 2.5) and references cited therein. For example, collapsed-stratum variance estimators generally will be conservative. In addition, for cases with moderate L, the variance estimators $V_{cs}^{*}(\hat{V}_B)$ and $V_{cs}^{*}(\hat{R}_{WV})$ may themselves have limited stability.

5. APPLICATION TO THE U.S. THIRD NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

5.1 Sample Design and Estimation Methods

The methods proposed in Sections 2 through 4 were applied to data from Phase I of the Third National Health and Nutrition Examination Survey (NHANES III). National Center for Health Statistics (1996) gives a general description of this survey, including special characteristics associated with Phase I (data collected between 1988 and 1991). For the present discussion, three aspects are of special interest. First, variance estimators were constructed on the basis of a collapsed design involving $L = 22$ strata (large groups of counties), with two primary sample units (generally individual counties) selected per stratum. Second, each selected PSU had a relatively large number of selected SSUs (generally groups of city blocks, or similar rural areas). The number of selected SSUs within each stratum ranged from 30 to 63, with a mean of 45.8.

Third, additional subsampling within each SSU led to selection of the survey elements (individual noninstitutionalized U.S. civilians). Each selected person was asked to respond to a health questionnaire and to participate in a detailed medical examination. Twelve of the resulting variables are listed in Table 1.

Standard weighted ratio estimates $\hat{\theta}$ were computed for the population means of each of the twelve variables listed in Table 1. The first two columns of Table 2 present the corresponding variance estimates $\hat{V}(\hat{\theta})$ and $\hat{V}_W$. As part of a larger study of the within-PSU variances V_{Wh} discussed in Jang and Eltinge (1996), there was considerable interest in the stability of the individual estimates $\hat{V}_{Wh}$. Since we had $n_h = 2$ for each stratum, the reasoning in Section 3.2 indicated that it was not feasible to examine the individual terms $\hat{d}_{Wh}$. Consequently, Section 5.2 will examine the pooled measure $\hat{d}_{W0}$ of the stability of the $\hat{V}_{Wh}$ and will also present some related simulation-based tests and diagnostic plots.

Table 1
Twelve NHANES III Variables

Variable name	Description
HAE2	Told by health professional that you had hypertension (indicator variable)
HAE7	Told by health professional that your blood cholesterol was high (indicator variable)
HAD1	Told by health professional that you had diabetes (indicator variable)
HAR3	Do you smoke cigarettes now?
BMPHT	Height
BMPWT	Weight
HDRESULT	HDL cholesterol
TCRESULT	Serum total cholesterol
LEAD	Blood lead, in micrograms per deciliter
log(LEAD)	Natural logarithm of blood lead
BP1K1	Systolic blood pressure
BP1K5	Diastolic blood pressure

Table 2
Variance Estimates and Stability Measures for Twelve NHANES III Variables

Variable name	$\hat{V}_W$	$\hat{V}(\hat{Y})$	$\hat{d}_{W0}$	$\hat{d}_{WF}$
HAE2	0.0000385	0.0000511	23.7	425.8
HAE7	0.0000821	0.000135	13.6	225.6
HAD1	0.00000956	0.00000749	8.8	160.6
HAR3	0.000122	0.000205	6.4	125.8
BMPHT	0.0223	0.0416	15.3	275.1
BMPWT	0.104	0.122	8.6	139.2
HDRESULT	0.0743	0.163	11.5	196.2
TCRESULT	0.590	0.860	21.2	353.9
LEAD	0.00388	0.00657	2.8	48.8
log(LEAD)	0.000211	0.000678	10.5	174.9
BP1K1	1.073	2.896	1.0	26.5
BP1K5	0.252	0.217	17.2	52.9

In addition, there was interest in the extent to which the variances of the $\hat{V}_{Wh}$ contributed to the variances of the pooled quantities $\hat{V}_B$ and $\hat{R}_{WV}$. Section 5.3 explores this question.

5.2 Within-PSU Variance Estimates and Associated Stability Measures

5.2.1 Comparison Across Variables

The final two columns of Table 2 report the degrees-of-freedom estimates $\hat{d}_{W0}$ and $\hat{d}_{WF}$ for the twelve NHANES III variables. Note especially that the stratum-level stability measures $\hat{d}_{W0}$ are relatively low, compared to the mean of 45.8 SSUs per stratum. For example, all of the variables have $\hat{d}_{W0}$ less than 24, and five (HAD1, HAR3, BMPWT, LEAD and BP1K1) have $\hat{d}_{W0}$ less than 10. Due to the interest in the $\hat{d}_{W0}$ described above, this led to two general questions.

(1) Are the observed $\hat{d}_{W0}$ consistent with the nominal degrees-of-freedom value d_{W0} that one would anticipate from the direct SSU counts $n_{h1} + n_{h2} - 2$?

(2) Conversely, are the observed $\hat{d}_{W0}$ consistent with distributional conditions that produce considerably smaller values of d_{W0}?

Standard large-sample-theory-based tests for (1) and (2) would have depended on eighth sample moments, and thus were inadvisable in the present case, due to the relatively small values of $L = 22$ and $n_h = 2$. Instead, the following simulation-based test was carried out.

5.2.2 Simulation-Based Interpretation of Stability Measures

This simulation work covers six cases involving different values of two terms. The first term, denoted d_{hi}, represents the degrees of freedom associated with the variance estimator $\hat{\sigma}^2_{2hi}$ in PSU (h, i). The second term, denoted R_{12}, is the ratio of the expressions $p_{hi}^{-2}\sigma^2_{2hi}$ in the first and second sample PSUs in stratum h.

In each of the six cases discussed below, independent pseudorandom variables g_{hi} were generated from a chi-square distribution on d_{hi} degrees of freedom for $h = 1, 2, \ldots, 22$ and $i = 1, 2$. Re-scaled variables $\hat{V}_{Whi} = d_{hi}^{-1} V_{Whi} g_{hi}$ were then computed, where V_{Whi} is a random variable equal to one with probability one-half and equal to R_{12} with probability one-half. The random variables g_{hi} and V_{Whi} are mutually independent. Finally, the sums $\hat{V}_{Wh} = \hat{V}_{Wh1} + \hat{V}_{Wh2}$ and the associated measures $\tilde{V}(\hat{V}_{Wh})$, $\tilde{V}(\hat{V}_W)$ and $\hat{d}_{W0}$ were computed. This was repeated 10,000 times.

Table 3 lists the values of d_{hi} and R_{12} covered in the six cases, and Table 4 lists the resulting simulated means, standard deviations and quantiles for $\hat{d}_{W0}$. When interpreting the results for these cases, note that randomness of the g_{hi} corresponds to the estimation error in the $\hat{\sigma}^2_{2hi}$ due to subsampling at the SSU and lower levels; and randomness of the V_{Whi} reflects the variability of the $p_{hi}^{-2}\sigma^2_{2hi}$ induced by sampling of PSUs within a given stratum.

Table 3
Cases Covered for the Simulated Quantiles

Cases	d	R_{12}
1	22	1
2	Obs. Dist.	1
3	5	1
4	22	9
5	Obs. Dist.	9
6	5	9

Case 1 uses $d_{hi} = 22$ and $R_{12} = 1$. Arguments from Section 3.3 show that the resulting $\hat{V}_{Wh}$ are distributed as constant multiples of a chi-square random variable with $d_{W0} = 44$ degrees of freedom. Thus, for Case 1, the choice of $d_{hi} = 22$ has led to simulated quantiles of $\hat{d}_{W0}$ that are approximately those that one would anticipate from the mean SSU count of 45.8 observed for Phase I of NHANES III, under the setting described in Section 3.4. Note that even in this idealized Case 1, the relative variability of the $\hat{d}_{W0}$ is fairly high.

Now compare the $\hat{d}_{W0}$ reported in Table 2 to the simulated quantiles from Case 1. All twelve of the observed $\hat{d}_{W0}$ fall below the 0.025 simulated quantile of 24.8; and ten of the twelve fall below the 0.005 quantile of 21.1. Thus, the $\hat{d}_{W0}$ observed for the NHANES III variables are not consistent with a nominal $d_{W0} = 44$ produced in the idealized setting covered by Case 1.

5.2.3 Simulation Under Alternative Conditions with Smaller d_{W0}

In general, the distribution of $\hat{d}_{W0}$ may deviate from that observed under the idealized Case 1 due to: (a) variability in the true SSU counts n_{hi}; (b) limited stability of the PSU-level estimates $\hat{\sigma}^2_{2hi}$; and (c) heterogeneity of the true PSU-level terms σ^2_{2hi}. Cases 2 through 6 cover the combined effects of these three factors.

Table 4
Simulated Quantiles for $\hat{d}_{W0}$

Cases	Mean	S.D.	$q_{.005}$	$q_{.01}$	$q_{.025}$	$q_{.05}$	$q_{.10}$	$q_{.25}$	$q_{.50}$	$q_{.75}$	$q_{.90}$	$q_{.95}$	$q_{.975}$	$q_{.99}$	$q_{.995}$
1	48.9	17.7	21.1	22.5	24.8	27.4	30.7	36.7	45.5	57.4	71.2	81.5	92.6	108.5	122.1
2	48.3	17.5	20.7	21.9	24.2	26.8	29.9	36.3	45.2	56.6	70.2	80.3	92.0	106.2	118.0
3	11.3	4.7	4.1	4.5	5.1	5.6	6.4	8.0	10.3	13.5	17.3	20.0	23.0	26.8	30.1
4	5.5	2.7	1.4	1.6	2.0	2.3	2.7	3.7	5.0	6.8	8.9	10.5	12.1	14.8	16.7
5	5.5	2.7	1.4	1.6	1.9	2.3	2.7	3.7	5.0	6.7	8.9	10.6	12.1	14.1	16.1
6	3.5	2.1	0.7	0.8	1.0	1.2	1.5	2.1	3.0	4.4	6.0	7.4	8.8	11.2	12.6

The design for Case 2 was identical to that for Case 1, except that the d_{hi} were random variables, selected with equal probabilities and with replacement from the 44 values $n_{hi} - 1$ corresponding to the 44 SSU counts n_{hi} in the original dataset. The resulting simulated quantiles of $\hat{d}_{W0}$ are similar to those for Case 1.

Case 3 uses $d_{hi} = 5$ and $R_{12} = 1$; the resulting $\hat{V}_{Wh}$ are distributed as constant multiples of chi-square random variables with $d_{W0} = 10$ degrees of freedom. The simulated quantiles for Case 3 were somewhat more consistent with the $\hat{d}_{W0}$ observed for the NHANES III dataset. For example, ten of the twelve variables have $\hat{d}_{W0}$ at or above the simulated 0.10 quantile of 6.4. However, two of the variables (lead and systolic blood pressure) had their $\hat{d}_{W0}$ below the simulated 0.005 quantile for Case 3.

Cases 4 through 6 cover more extreme cases of instability, induced by use of the scale factor $R_{12} = 9$. A scale factor different from one introduces a component of variability associated with sampling of PSUs with unequal σ^2_{2hi}, and causes the $\hat{V}_{Wh}$ to have distributions outside of the rescaled chi-square family. Cases 4 through 6 use the same d_{hi} values used in Cases 1 through 3, respectively. The smallest observed NHANES III $\hat{d}_{W0}$ values are somewhat more consistent with the simulated quantiles for Cases 4 through 6, although the $\hat{d}_{W0} = 1.0$ for systolic blood pressure still falls below the simulated 0.005 quantile for Cases 4 and 5, and is approximately equal to the simulated 0.025 quantile for Case 6.

In addition, note that the three largest observed $\hat{d}_{W0}$ values (for the hypertension indicator, the total cholesterol measure, and diastolic blood pressure) fall above the simulated upper 0.995 quantiles for each of cases 4 through 6. This, in conjunction with the abovementioned results for Cases 1 through 3, indicates that the twelve observed $\hat{d}_{W0}$ are consistent with settings that produce substantially different true d_{W0} values for different variables.

Taken together, these simulation results suggest that for the twelve NHANES III variables examined, the stability of $\hat{V}_{Wh}$ may be substantially worse than one would anticipate from a simple count of SSUs within each stratum; and that the true stability measures d_{W0} may vary substantially from one variable to the next.

5.2.4 Diagnostic Plots

In a purely numerical sense, $\hat{d}_{W0}$ depends on the magnitudes of the $\tilde{V}(\hat{V}_{Wh})$ relative to the terms $2\hat{V}^2_{Wh}$. Consequently, diagnostic plots of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ against $\hat{V}_{Wh}$ are useful in the identification of specific patterns and "problem strata" that lead to unusually high or low $\hat{d}_{W0}$.

Figures 1 through 3 give plots for the variables HAE2 (diagnosed hypertension), log(blood lead), and blood lead, respectively. Each plot was constructed with horizontal and vertical axes on the same scale. The plot for HAE2 has the bulk of its points well below a line with slope = 1 and intercept = 0. In addition, the values of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ that are large in an absolute sense are still substantially less than the corresponding $\hat{V}_{Wh}$. This is consistent with the relatively large degrees-of-freedom value $\hat{d}_{W0} = 23.7$. The plot for log(blood lead) shows a somewhat greater concentration of points near the line with slope = 1 and intercept = 0, which is consistent with the somewhat smaller value $\hat{d}_{W0} = 10.5$.

The plot for blood lead shows one apparent outlier: the largest value of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ is approximately equal to the corresponding $\hat{V}_{Wh}$. For this stratum, we examined the terms $\hat{V}_{Wh}$ and $p^{-2}_{hi}\hat{\sigma}^2_{2hi}$ for unusual patterns, *e.g.*, extreme individual

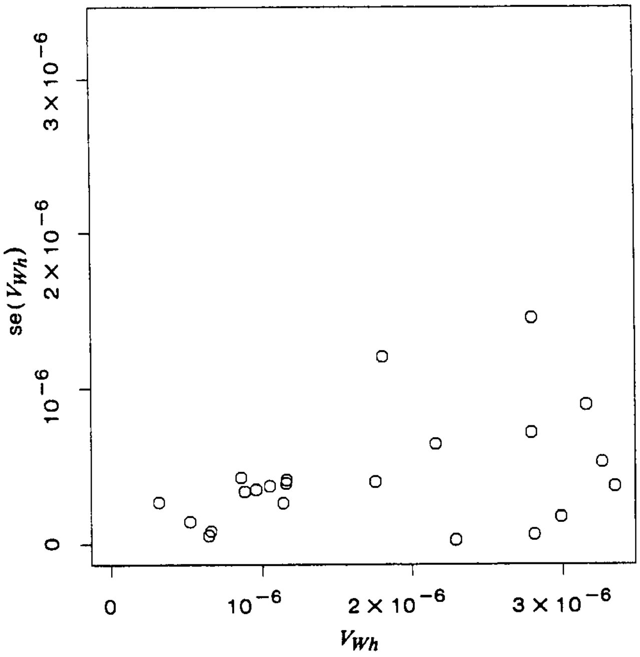

Figure 1. Plot of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ against $\hat{V}_{Wh}$ for HAE2

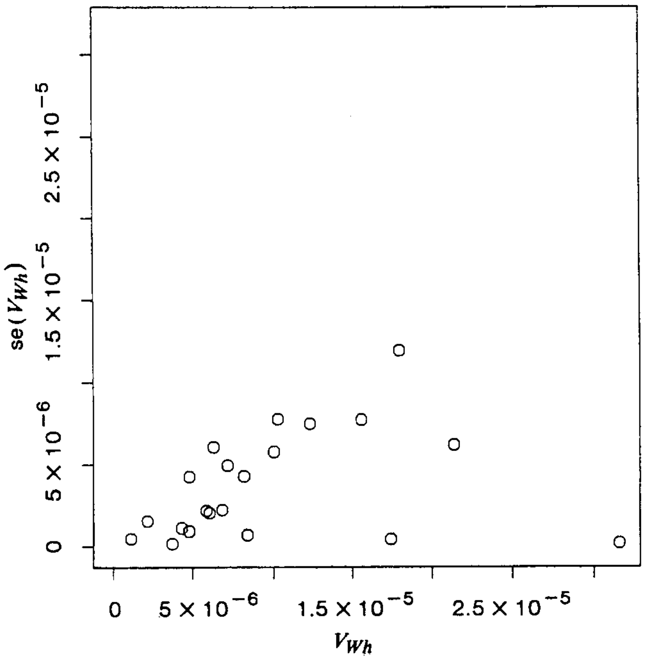

Figure 2. Plot of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ against $\hat{V}_{Wh}$ for log (blood lead)

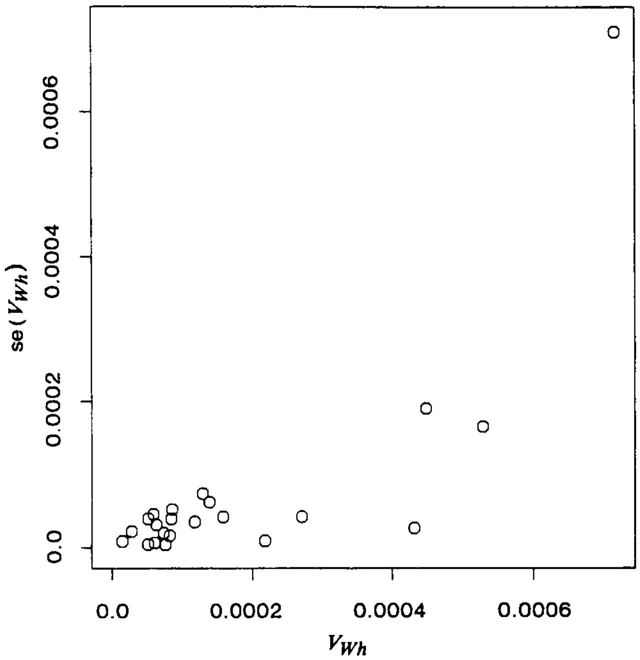

Figure 3. Plot of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ against $\hat{V}_{Wh}$ for blood lead

values or extreme element-level weights. Here, one of the two associated $p_{hi}^{-2}\hat{\sigma}_{2hi}^{2}$ values was approximately equal to zero and the other was the largest of all the PSU-level terms $p_{hi}^{-2}\hat{\sigma}_{2hi}^{2}$. In addition, the stratum in question had the largest $\hat{V}_h$ value. However, this stratum did not display outlying values of $\tilde{V}(\hat{V}_{Wh})^{1/2}$ and $\hat{V}_h$ for other related variables, *e.g.*, log (blood lead). Thus, the unusual pattern observed for blood lead may be attributable to a few very high observed values for the blood lead variable, rather than to the sample design or weighting as such. Within this context, note that at the population level in the U.S., lead measurements tend to have a roughly lognormal distribution, and high lead measurements show some tendency to be clustered together due to environmental factors.

5.3 Between-PSU Variance Estimates and the Variance Ratio $\hat{R}_{WV}$

Table 5 presents the estimates $\hat{V}_B$ and $\hat{R}_{WV}$, and associated standard errors, for the twelve NHANES III variables. Of special interest are the columns labeled $\tilde{V}(\hat{V}_B)^{-1}\tilde{V}(\hat{V}_W)$, the proportion of the variance estimate $\tilde{V}(\hat{V}_B)$ that is attributable to the within-PSU variance term; and $\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^{2}\tilde{V}(\hat{V}_W)$, the corresponding proportion for $\hat{R}_{WV}$. Relatively large values for these proportions indicate that $\tilde{V}(\hat{V}_W)$ makes a substantial contribution to $\tilde{V}(\hat{V}_B)$ and $\tilde{V}(\hat{R}_{WV})$ for the variables in question.

Note that the proportion $\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^{2}\tilde{V}(\hat{V}_W)$ is greater than or equal to 0.3 for blood lead, BP1K1 (systolic blood pressure) and BP1K5 (diastolic blood pressure). For blood lead and BP1K1, the large proportions arise primarily because of the relatively large value of $\tilde{V}(\hat{V}_W)$. For BP1K5, $\tilde{V}(\hat{V}_W)$ is not as large on a relative scale, but the proportion $\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^{2}\tilde{V}(\hat{V}_W)$ is still large because $\hat{V}_W$ is not small relative to $\tilde{V}(\hat{Y})$. For all three variables, the relatively large values of $\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^{2}\tilde{V}(\hat{V}_W)$ indicate that it is important to account for the variance $V(\hat{V}_W)$ when one considers the stability of $\hat{R}_{WV}$. For BP1K5, a similar comment applies to the effect of $V(\hat{V}_W)$ on the stability of $\hat{V}_B$.

Table 5
Estimates of $\hat{V}_B$ and $\hat{R}_{WV}$ for Twelve NHANES III Variables with Associated Standard Errors and Relative Within-PSU Contributions

Variable name	$\hat{V}_B$	$\text{se}(\hat{V}_B)$	$\tilde{V}(\hat{V}_B)^{-1}\tilde{V}(\hat{V}_W)$
HAE2	0.0000126	0.0000188	0.020
HAE7	0.0000532	0.0000445	0.030
HAD1	-0.00000208	0.00000246	0.186
HAR3	0.0000825	0.0000703	0.047
BMPHT	0.0193	0.0114	0.027
BMPWT	0.0174	0.0400	0.096
HDRESULT	0.0887	0.0744	0.010
TCRESULT	0.270	0.253	0.031
LEAD	0.00269	0.00188	0.168
log(LEAD)	0.000468	0.000205	0.012
BP1K1	1.823	0.997	0.081
BP1K5	-0.0351	0.0793	0.367

	$\hat{R}_{WV}$	$\text{se}(\hat{R}_{WV})$	$\tilde{V}(\hat{R}_{WV})^{-1}\hat{V}_W^{-2}\hat{R}_{WV}^{2}\tilde{V}(\hat{V}_W)$
HAE2	1.327	0.491	0.034
HAE7	1.648	0.556	0.077
HAD1	0.783	0.247	0.123
HAR3	1.676	0.600	0.122
BMPHT	1.864	0.530	0.089
BMPWT	1.168	0.391	0.126
HDRESULT	2.193	1.020	0.047
TCRESULT	1.458	0.436	0.063
LEAD	1.694	0.555	0.367
log(LEAD)	3.221	1.025	0.112
BP1K1	2.699	1.142	0.391
BP1K5	0.861	0.300	0.300

6. DISCUSSION

This paper has presented three main ideas. First, due to the role that estimated within-PSU variances $\hat{V}_{Wh}$ play in survey design and analysis, it is important to account for the sampling error encountered in estimation of V_{Wh}. Second, standard design-based estimation methods lead to relatively simple estimators of the design variance of $\hat{V}_{Wh}$. In general, interpretation of these stability measures requires some caution. However, they can provide useful diagnostics for the identification of variables for which the instability of $\hat{V}_{Wh}$ is especially problematic, or has an especially pronounced effect on the variance of related quantities like $\hat{V}_B$ and $\hat{R}_{WV}$. Third, the application to the U.S. Third National Health and Nutrition Examination Survey (NHANES III), and associated simulation work, indicated the following.

(i) For different sets of variables, the observed stability measures $\hat{d}_{W0}$ are consistent with substantially different sets of stability conditions.

(ii) For some variables, the estimators $\hat{V}_{Wh}$ are considerably less stable than one would anticipate from a direct count of secondary sample units.

(iii) For some variables, the estimated variance of $\hat{V}_{Wh}$ makes a substantial contribution to the estimated variances of the estimated between-PSU variance $\hat{V}_B$ and the variance ratio $\hat{R}_{WV}$.

ACKNOWLEDGEMENTS

We thank Van Parsons, Cliff Johnson and other NCHS statisticians for providing access to the NHANES III dataset; and for sharing a wealth of information regarding the NHANES III - Phase I project. We also thank Van Parsons and two anonymous referees for helpful comments on earlier versions of this paper. This research was supported in part by the National Center for Health Statistics. The views expressed in this paper are those of the authors and do not necessarily represent the policies of the National Center for Health Statistics.

REFERENCES

FULLER, W.A. (1984). Least squares and related analyses for complex survey design. *Survey Methodology*, 10, 97-118.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory, Volume I: Methods and Applications*. New York: John Wiley.

HERZOG, T.N., and SCHEUREN, F.J. (1976). Dallying with some CPS design effects for proportions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-401.

JANG, D.S., and ELTINGE, J.L. (1996). Use of Within-PSU Variances and Errors-in-Variables Regression to Assess the Stability of a Standard Design-Based Variance Estimator. Unpublished manuscript, Department of Statistics, Texas A&M University.

KORN, E.I., and GRAUBARD, B.G. (1995). Analysis of large health surveys: accounting for the sampling design. *Journal of the Royal Statistical Society*, Series A, 158, 263-295.

NATIONAL CENTER FOR HEALTH STATISTICS (1996). National Health and Nutrition Examination Survey III Report (in press). National Center for Health Statistics, Hyattsville, MD.

RUST, K., and KALTON, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics*, 3, 69-81.

SATTERTHWAITE, F.E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.

SATTERTHWAITE, F.E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110-114.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.