DESIGN EFFECTS IN SCHOOL SURVEYS

William H. Robb and Ronaldo Iachan
ORC Macro, Burlington, Vermont 05401

KEY WORDS: design effects, generalized variance functions, school surveys, weighting effects

School surveys typically involve complex multistage stratified sampling designs. In addition, survey objectives often require over-sampling of minority groups, and the reporting of hundreds of estimates computed for many different subgroups of interest. Another feature of these surveys that limits the ability to control the precision for the overall population and key subgroups simultaneously is that intact classes are selected as ultimate clusters; thus, the designer cannot determine the ethnicity of individual students nor the ethnic composition of classes.

ORC Macro statisticians have developed the design and weighting procedures for numerous school surveys at the national and state levels. While this analysis is focused on the national Youth Risk Behavior Survey (YRBS), we have encountered similar problems in surveys such as the National Youth Tobacco Survey (NYTS) and the Health Behavior in School Children (HBSC) studies.

In surveys that are repeated periodically—or surveys that use similar designs—detailed investigations of the relative precision of these estimates can improve the survey design and weighting procedures. These investigations may be focused on the design effect (DEFF), defined as the variance under the actual sampling design divided by the variance under a simple random sample of the same size. The DEFF may vary substantially across the different types of estimates (e.g., survey items) computed in the survey; however, similar patterns may be observed for similar estimates, or subgroups.

It is customary to decompose the DEFF into a component due to unequal weighting and another component that reflects clustering effects (primarily). This investigation represents our first attempts to assess these effects for different groups of items. To gain additional insight into the pattern of variability by item group and by subgroup of students, the analysis also looks at Generalized Variance Functions (GVF).

The Youth Risk Behavior Survey (YRBS)

The YRBS is a school-based survey designed to provide national estimates of the prevalence of risk behaviors conducted by the Division of School and Adolescent Health (DASH), Centers for Disease Control and Prevention (CDC). Items on the survey assess engagement in a variety of risk behaviors that include drug use (including tobacco and alcohol use), sexual activity, personal safety, nutrition, school violence, personal safety and physical activity. The YRBS is designed to provide estimates for the population of ninth through twelfth graders in public and private schools, by gender, by age or grade, and by grade and gender for all youths, for African-American youths, and for Hispanic youths. The YRBS is conducted by ORC Macro under contract to CDC-DASH on a two year cycle.

For the cycle of the YRBS examined in this paper, 57 primary sampling units (PSUs) were selected from within 16 strata. Within PSUs, at least three schools per cluster spanning grades 9 through 12 were drawn. Fifteen additional schools were selected in a subsample of PSUs to represent very small schools. In addition, schools that did not span the entire grade range 9 through 12 were combined prior to sampling so that every sampling unit spanned the grade range 9-12. The sample included 199 schools, of which 150 participated. One class was selected per school per grade, except for schools with the largest concentrations of minority students where two classes were selected per grade. The 2001 YRBS survey yielded 13,627 responses.

The sample design employed a three-stage cluster sample stratified by racial/ethnic concentration, geographic location and MSA status. Within each stratum, a sample of primary sampling units (PSUs)-- a county or a group of counties--was chosen from which a probabilistic selection of schools and students was subsequently made.

Three strategies were employed to achieve over-sampling of African-Americans and Hispanics: a) larger sampling rates were used in high-Hispanic and high-African-American strata; b) a modified measure of size was employed that increased the probability of selection of schools with disproportionately high minority enrollments; and c) two classes per grade (rather than one) were selected in high-minority schools.

Sampling was with probabilities proportional to size (PPS) at first and second stages for selecting PSUs and schools. The modified measure of size used was a weighted linear combination of Hispanic, African American and Other enrollments. To achieve a nearly equal probability of selection for students, an approximately constant number of students is selected at the final stage in clusters of classrooms.

The weighting process started with the computation of a basic sampling weight as the inverse of the probability of selection for each student. The weights were then adjusted for non-response at the student and school level. Following this adjustment, weights were trimmed and post-stratified to match population counts obtained from the

sampling frame. More details on the sampling and weighting procedures can be found in Iachan and Robb (2002).

Computation of Design Effects

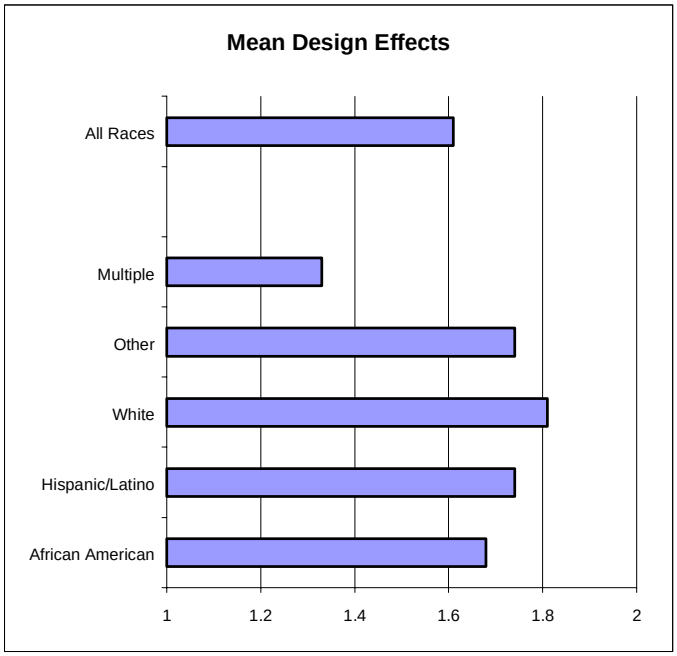
For this study, SUDAAN was used to compute the following estimates.

- Percentage of 'response of interest' for each item
- Design effect for percentage estimates
- Standard error for percentage estimates
- Totals for 'response of interest' for each item
- Standard error for total estimates

The 'response of interest' was constructed by the CDC-DASH staff, and coded each multi-response item into yes/no set of responses. For example, for the item asking "During the past 30 days, on how many days did you carry a weapon such as a gun, knife, or club?' the response of interest was defined as 1 for any response category indicating that a weapon was carried on one or more days.

Estimates were computed by race/ethnicity and by gender; estimates were grouped into 12 categories based on the type of item. Mean design effects and standard errors were computed across item categories. Standard errors averaged across all estimates by race ranged from 0.88 for estimates among Whites to 2.84 for estimates among those of multiple races, reflecting the varying group sample sizes.

Figure 1 depicts mean design effects for estimates by race. Note that as these design effects are computed directly from the estimates, they combine clustering effects and unequal weighting effects.



The portion of the design effect that can be attributed to unequal weighting effects can be written as

$$deffw = 1 + \left(cv(w)^2\right)$$

where cv(w) is the coefficient of variation of the final weights. The design effect due to weighting was computed for the same set of race/ethnic groups.

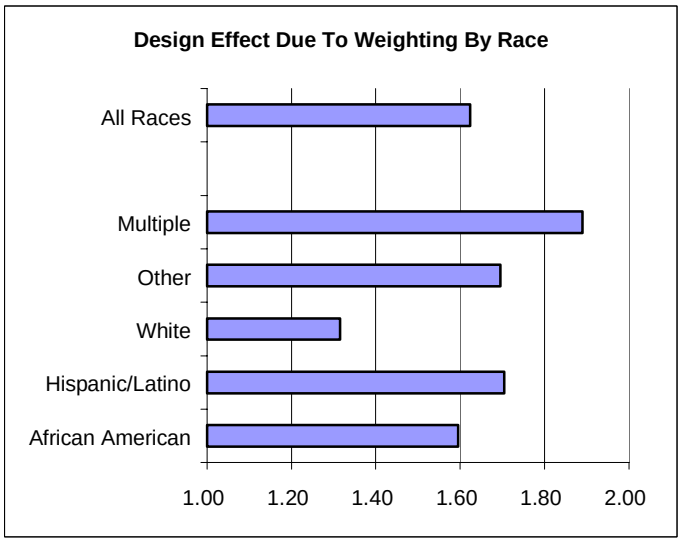
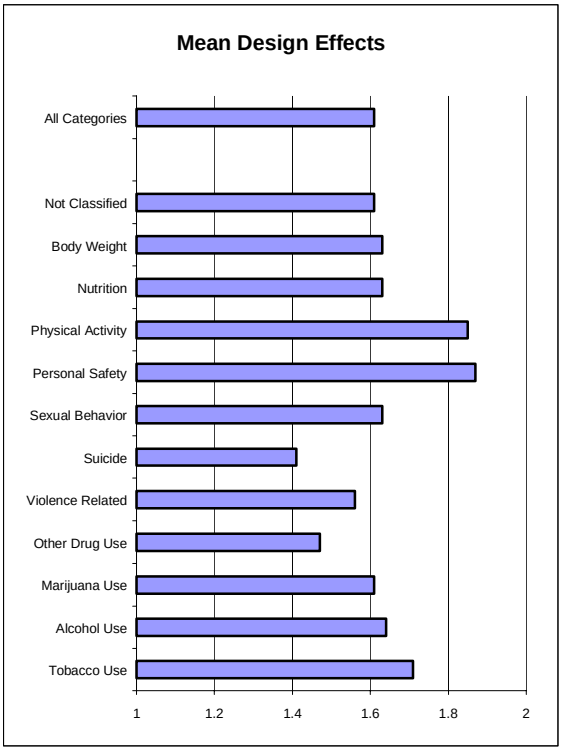
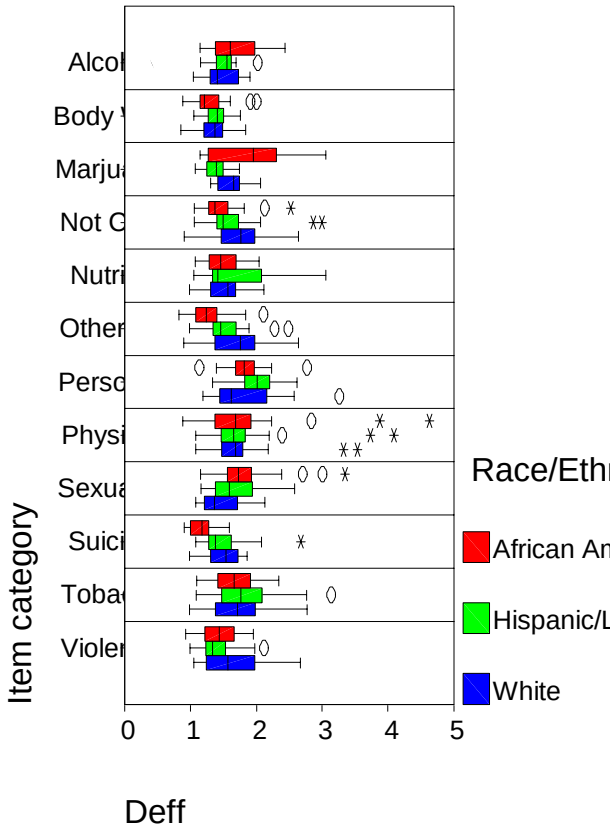


Figure 2 shows that unequal weighting effects are highest for the multiple race group where respondents come from schools (over)sampled at widely varying rates. Weights are also highly variable for Blacks and Hispanics, groups that are heavily represented in the over-sampled schools and PSUs. Black and Hispanic respondents, however, may also be present in other schools selected at much lower sampling rates. Note that while unequal weighting effects are very low for Whites, overall DEFFs shown in Figure 1 are large for Whites on average reflecting that large clustering effects are present for this group as well.

Figure 3 presents mean design effects by item category. There is considerable variation in the design effects across different item groups. This variability is most likely due to variations in the degree of intra-cluster correlation among the items. Within the item category "Physical Activity", for example, items such as "days attended PE class" will have a high degree of intra-school correlation, and therefore tend to have higher design effects. Similarly, students tend to feel safer (or less safe) within a same school, so the Personal Safety item category also contains items with large DEFFs due to large intra-cluster correlations. Other behaviors may cluster among groups of schools – that is, at the PSU level – giving much the same effect.



Having seen an effect due to race, as well as an effect due to item grouping, we examined possible interactions between the racial grouping and item grouping effects. Figure 4 presents box-and-whisker plots of design effects by race and item grouping. (In this chart, the extreme values are represented by circles, and outliers by stars.) The outliers appear in the physical activity group for items that are more dependent on school policy than on student behavior, such as days of PE class.



For some items, such as violence behaviors, there appears to be little difference in the design effects across racial groupings. However, the mean design effect across marijuana use estimates was higher for the African-American than for the Hispanic population. These estimates are subject to a combination of higher weighting effects and intra-cluster correlations.

The presence of outliers, and the general variability in DEFFs, suggest that caution is recommended in the reporting and use of average DEFFs, a practice that is often followed. To gain additional insight into patterns of variability, the modeling of Generalized Variance Functions (GVF) was also pursued.

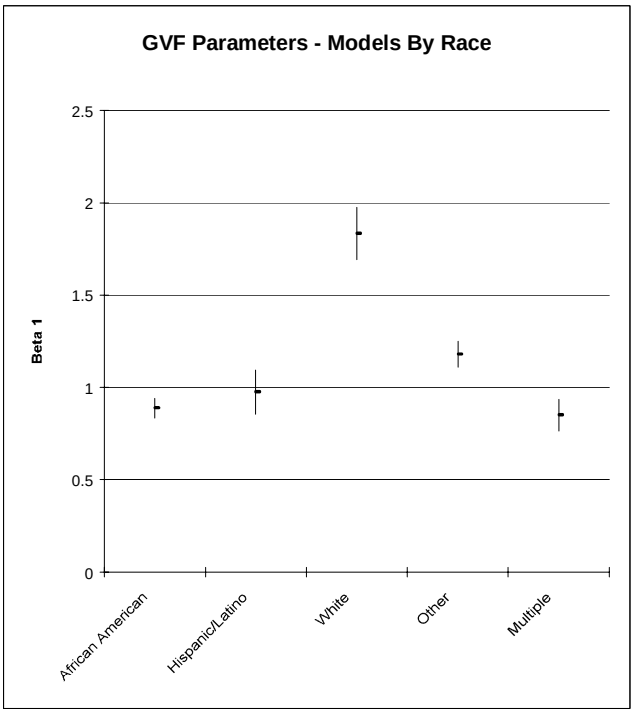
Generalized Variance Functions

Methods for modeling (relative) variances and standard errors have been explored from practical and conceptual perspectives (e.g., Wolter, 1985; Bieler and Williams, 1990). These models are typically based on modeling the Relative Variance,

$$RV(\hat{Y}) = \frac{Var(\hat{Y})}{\hat{Y}^2}$$

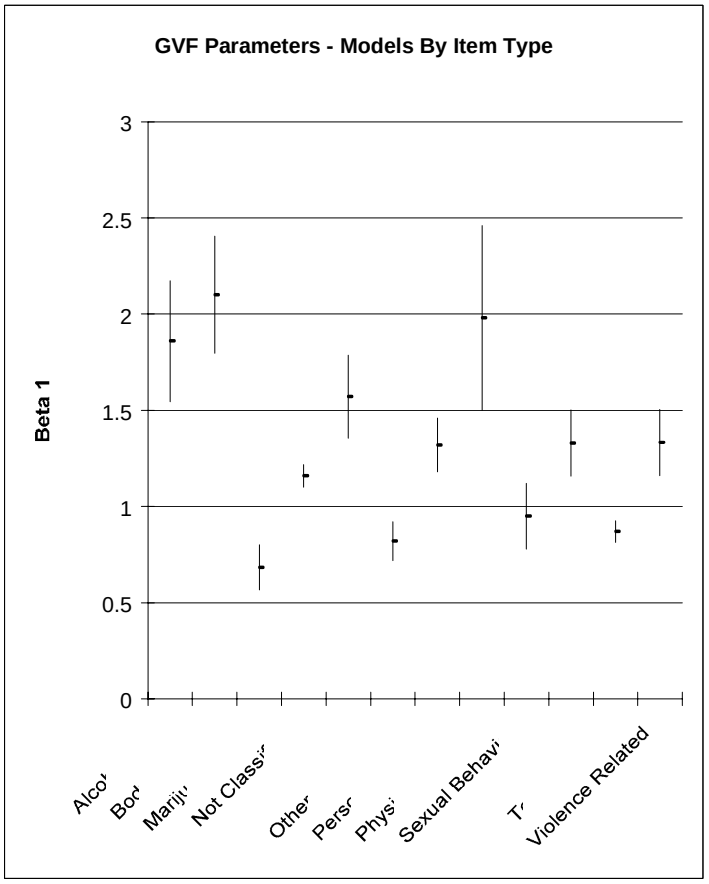
in functions of the following type:

$$RV(\hat{Y}) = \beta_0 + \frac{\beta_1}{\hat{Y}}$$



Figures 5 and 6 present GVF models fit for the various estimates grouped by race and by item category. These figures show the coefficient (Beta) fit in the models together with the associated confidence intervals, CIs (for Beta). Generally, the models had good fit (data not shown) although it may be noted that the CIs are wide for certain item categories, primarily Physical Activity but also Body Weight and Alcohol Use.

The use of such models may also facilitate the reporting of hundreds of (variance) estimates needed in the YRBS.



References

Bieler, G.S. and Williams, R.L. (1990). Generalized Standard Error Models for Proportions in Complex Design Surveys", Proceedings of the Survey Research Methods of the American Statistical Association, 1990, pp. 272-277.

Iachan, R. and Robb, W. H. (2002) *Sampling and Weighting Report for the 2001 Youth Risk Behavior Survey*.

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