Chapter VI
Estimating components of design effects for use in sample design

Graham Kalton
Westat
Rockville, Maryland
United States of America

J. Michael Brick
Westat
Rockville, Maryland
United States of America

Thanh Lê
Westat
Rockville, Maryland
United States of America

Abstract

The design effect - the ratio of the variance of a statistic with a complex sample design to the variance of that statistic with a simple random sample or an unrestricted sample of the same size - is a valuable tool for sample design. However, a design effect found in one survey should not be automatically adopted for use in the design of another survey. A design effect represents the combined effect of a number of components such as stratification, clustering, unequal selection probabilities, and weighting adjustments for non-response and non-coverage. Rather than simply importing an overall design effect from a previous survey, careful consideration should be given to the various components involved. The present chapter reviews the design effects due to individual components, and then describes models that may be used to combine these component design effects into an overall design effect. From the components, the sample designer can construct estimates of overall design effects for alternative sample designs and then use these estimates to guide the choice of an efficient sample design for the survey being planned.

Key terms: stratification, clustering, weighting, intra-class correlation coefficient.

A. Introduction

1. As can be seen from other chapters in the present publication, national household surveys in developing and transition countries employ complex sample designs, including multistage sampling, stratification, and frequently unequal selection probabilities. A consequence of the use of a complex sample design is that the sampling errors of the survey estimates cannot be computed using the formulae found in standard statistical texts. Those formulae are based on the assumption that the variables observed are independently and identically distributed (*iid*) random variables. That assumption does not hold for observations selected by complex sample designs, and hence a different approach to estimating the sampling errors of survey estimates is needed.

2. Variances of survey estimates from complex sample designs may be estimated by some form of replication method, such as jackknife repeated replication or balanced repeated replication, or by a Taylor series linearization method [see, for example Wolter (1985); Rust (1985); Verma (1993); Lehtonen and Pahkinen (1994); Rust and Rao (1996)]. A number of specialized computer programs are available for performing the computations [see reviews of many of them by Lepkowski and Bowles (1996), also available at http://www.fas.harvard.edu/~stats/survey-soft/iass.html; and the summary of survey analysis software, prepared by the Survey Research Methods Section of the American Statistical Association, available at http://www.fas.harvard.edu/~stats/survey-soft/survey-soft html]. When variances are computed in a manner that takes account of the complex sample design, the resulting variance estimates are different from those that would be obtained from the application of the standard formulae for *iid* variables. In many cases, the variances associated with a complex design are larger -- often appreciatively larger -- than those obtained from standard formulae.

3. The variance formulae found in standard statistical texts are applicable for one form of sample design, namely, unrestricted sampling (also known as simple random sampling with replacement). With this design, units in the survey population are selected independently and with equal probability. The units are sampled with replacement, implying that a unit may appear more than once in the sample. Suppose that an unrestricted sample of size n yields values $y_1, y_2, ..., y_n$ for variable y. The variance of the sample mean $\bar{y} = \Sigma y_i / n$ is

$$V_u(\bar{y}) = \sigma^2 / n \tag{1}$$

where $\sigma^2 = \sum^N (Y_i - \bar{Y})^2 / N$ is the element variance of the N y-values in the population $(Y_1, Y_2, ..., Y_N)$ and $\bar{Y} = \Sigma Y_i / N$. This variance may be estimated from the sample by

$$v_u(\bar{y}) = s^2 / n \tag{2}$$

where $s^2 = \sum^n (y_i - \bar{y})^2 / (n-1)$. The same formulae are to be found in standard statistical texts.

4. As a rule, survey samples are selected without, rather than with, replacement because the survey estimates are more precise (that is to say, they have lower variances) when units can be included in the sample only once. With simple random sampling without replacement, generally known simply as simple random sampling or SRS, units are selected with equal probability, and all possible sets of n distinct units from the population of N units are equally likely to constitute the sample. With a SRS of size n, the variance and variance estimate for the sample mean $\bar{y} = \Sigma y_i / n$ are given by

$$V_0(\bar{y}) = (1-f)S^2 / n \tag{3}$$

and

$$v_0(\bar{y}) = (1-f)s^2 / n \tag{4}$$

where $f = n/N$ is the sampling fraction, $S^2 = \sum^N (Y_i - \bar{Y})^2 /(N-1)$, and $s^2 = \sum^n (y_i - \bar{y})^2 /(n-1)$. When N is large, as is generally the case in survey research, σ^2 and S^2 are approximately equal. Thus, the main difference between the variance for the mean for unrestricted sampling in equation (1) and that for SRS in (3) is the factor $(1-f)$, known as the finite population correction (fpc). In most practical situations, the sampling fraction n/N is small, and can be treated as 0. When this applies, the fpc term in (3) and (4) is approximately 1, and the distinction between sampling with and without replacement can be ignored.

5. The variance formulae given above are not applicable for complex sample designs, but they do serve as useful benchmarks of comparison for the variances of estimates from complex designs. Kish (1965) coined the term "design effect" to denote the ratio of the variance of any estimate, say, z, obtained from a complex design to the variance of z that would apply with a SRS or unrestricted sample of the same size.[18] Note that the design effect relates to a specific survey estimate z, and will be different for different estimates in a given survey. Also note that z can be any estimate of interest, for instance, a mean, proportion, total, or regression coefficient.

6. The design effect depends both on the form of complex sample design employed and on the survey estimate under consideration. To incorporate both these characteristics, we employ the notation $D^2(z)$ for the design effect of the estimate z, where

[18] More precisely, Kish (1982) defined $Deff$ as this ratio with a denominator of the SRS variance, and $Deft^2$ as the ratio with a denominator of the unrestricted sample variance. The difference between *Deff* and $Deft^2$ is based on whether the fpc term $(1-f)$ is included or not. Since that term has a negligible effect in most national household surveys, the distinction between *Deff* and $Deft^2$ is rarely of practical significance, and will therefore be ignored in the remainder of this chapter. Throughout, we assume that the fpc term can be ignored. See also Kish (1995).

Skinner defined a different but related concept, the mis-specification effect or *meff*, which he argues, is more appropriate for use in analysing survey data (see, for example, Skinner, Holt and Smith (1989), chap. 2). Since this chapter is concerned with sample design rather than analysis, that concept will not discussed here.

$$D^2(z) = \frac{\text{Variance of } z \text{ with the complex design}}{\text{Variance of } z \text{ with an unrestricted sample of the same size}} = \frac{V_c(z)}{V_u(z)} \qquad (5)$$

The squared term in this notation is employed to enable the use of $D(z)$ as the square root of the design effect. A simple notation for $D(z)$ is useful since it represents the multiplier that should be applied to the standard error of z under an unrestricted sample design to give its standard error under the complex design as in, for instance, the calculation of a confidence interval.

7. A useful concept directly related to the design effect is "effective sample size", denoted here as n_{eff}. The effective sample size is the size of an unrestricted sample that would yield the same level of precision for the survey estimate as that attained by the complex design. Thus, the effective sample size is given by

$$n_{eff} = n / D^2(z) \qquad (6)$$

8. The definition of $D^2(z)$ given above is for theoretical work where the true variances $V_c(z)$ and $V_0(z)$ are known. In practical applications, these variances are estimated from the sample, and $D^2(z)$ is then estimated by $d^2(z)$. Thus,

$$d^2(z) = \frac{v_c(z)}{v_u(z)} \qquad (7)$$

where $v_c(z)$ is estimated using a procedure appropriate for the complex design and $v_u(z)$ is estimated using a formula for unrestricted sampling with unknown parameters estimated from the sample. Thus, for example, in the case of the sample mean

$$v_u(z) = s^2 / n \qquad (8)$$

and, for large samples, s^2 may be estimated by

$$\frac{\sum w_i (y_i - \bar{y})^2}{\sum w_i}$$

where y_i and w_i are the *y*-value and the weight of sampled unit *i* and $\bar{y} = \sum w_i y_i / \sum w_i$ is the weighted estimate of the population mean. In the case of a sample proportion *p*, for large *n*

$$v_u(p) = \frac{p(1-p)}{n-1}$$

or

$$v_u(p) = \frac{p(1-p)}{n}$$

where p is the weighted estimate of the population proportion.

9. In defining design effects and estimated design effects, there is one further issue that needs to be addressed. Many surveys employ sample designs with unequal selection probabilities and when this is so, subgroups may be represented disproportionately in the sample. For example, in a national household survey, 50 per cent of a sample of 2,000 households may be selected from urban areas and 50 per cent from rural areas, whereas only 30 per cent of the households in the population are in urban areas. Consider the design effect for an estimated mean for, say, urban households. The denominator from (8) is s^2/n. The question is how n is to be computed. One approach is to use the actual urban sample size, 1,000 in this case. An alternative is to use the expected sample size in urban areas for a SRS of n = 2,000, which here is $0.3 \times 2000 = 600$. The first of these approaches, which conditions on the actual size of 1,000, is the one that is most commonly used, and it is the approach that will be used in this chapter. However, the option to compute design effects based on the second approach is available in some variance estimation programs. Since the two approaches can produce markedly different values, it is important to be aware of the distinction between them and to select the appropriate option.

10. The concept of design effect has proved to be a valuable tool in the design of complex samples. Complex designs involve a combination of a number of design components, such as stratification, multistage sampling, and selection with unequal probabilities. The analysis of the design effects for each of these components individually sheds useful light on their effects on the precision of survey estimates, and thus helps guide the development of efficient sample designs. We review the design effects for individual components in section B. In designing a complex sample, it is useful to construct models that predict the overall design effects arising from a combination of components. We briefly review these models in section C. We provide an illustrative hypothetical example of the use of design effects for sample design in section D, and conclude with some general observations in section E.

B. Components of design effects

11. The present section considers the design effects resulting from the following components of a complex sample design: proportionate and disproportionate stratification; clustering; unequal selection probabilities; and sample weighting adjustments for non-response, and population weighting adjustments for non-coverage and for improved precision. These various components are examined separately in this section; their joint effects are discussed in section C. The main statistic considered is an estimate of a population mean $\bar{Y}$ (for example, mean income). Since a population proportion P (for example, the proportion of the population living in poverty) is in fact a special case of an arithmetic mean, the treatment covers a proportion also. Proportions are probably the most widely used statistics in survey reports, and they will therefore be discussed separately when appropriate. Many survey results relate to subgroups of the total

population, such as women aged 15 to 44, or persons living in rural areas. The effects of weighting and clustering on the design effects of subgroup estimates will therefore be discussed.

1. Stratification

12. We start by considering the design effect for the sample mean in a stratified single-stage sample with simple random sampling within strata. The stratified sample mean is given by

$$\bar{y}_{st} = \Sigma_h \frac{N_h}{N} \Sigma_i \frac{y_{hi}}{n_h} = \Sigma_h W_h \bar{y}_h$$

where n_h is the size of the sample selected from the N_h units in stratum h, $N = \Sigma N_h$ is the population size, $W_h = N_h / N$ is the proportion of the population in stratum h, y_{hi} is the value for sampled unit i in stratum h, and $\bar{y}_h = \Sigma_i y_{hi} / n_h$ is the sample mean in stratum h. In practice, $\bar{y}_{st}$ is computed as a weighted estimate, where each sampled unit is assigned a base weight that is the inverse of its selection probability (ignoring for the moment sample and population weighting adjustments). Here each unit in stratum h has a selection probability of n_h / N_h and hence a base weight of $w_{hi} = w_h = N_h / n_h$. Thus, $\bar{y}_{st}$ may be expressed as

$$\bar{y}_{st} = \frac{\Sigma_h \Sigma_i w_{hi} y_{hi}}{\Sigma_h \Sigma_i w_{hi}} = \frac{\Sigma_h \Sigma_i w_h y_{hi}}{\Sigma_h n_h w_h} \tag{9}$$

Assuming that the finite population correction can be ignored, the variance of the stratified mean is given by

$$V\left(\bar{y}_{st}\right) = \Sigma_h \frac{W_h^2 S_h^2}{n_h} \tag{10}$$

where $S_h^2 = \Sigma_i (Y_{hi} - \bar{Y}_h)^2 / (N_h - 1)$ is the population unit variance within stratum *h*.

13. The magnitude of $V(\bar{y}_{st})$ depends upon the way the sample is distributed across the strata. In the common case where a proportionate allocation is used, so that the sample size in a stratum is proportional to the population size in that stratum, the weights for all sampled units are the same. The stratified mean reduces to the simple unweighted mean $\bar{y}_{prop} = \Sigma\Sigma y_{hi} / n$, where $n = \Sigma n_h$ is the overall sample size, and its variance reduces to

$$V(\bar{y}_{prop}) = \frac{\Sigma W_h S_h^2}{n} = \frac{S_w^2}{n} \tag{11}$$

where S_w^2 denotes the average within-stratum unit variance. The design effect for $\bar{y}_{prop}$ for a proportionate stratified sample is then obtained using the variance of the mean for a simple random sample from equation (3), ignoring the fpc term, and with the definition of the design effect in equation (5) as

$$D^2(\bar{y}_{prop}) = \frac{S_w^2}{S^2} \tag{12}$$

Since the average within-stratum unit variance is no larger than the overall unit variance (provided that the values of N_h are large), the design effect for the mean of a proportionate sample is no greater than 1. Thus, proportionate stratification cannot lead to a loss in precision, and generally leads to some gain in precision. A gain in precision occurs when the strata means $\bar{Y}_h$ differ: the larger the variation between the means, the greater the gain.

14. In many surveys, a disproportionate stratified sample is needed to enable the survey to provide estimates for particular domains. For example, an objective of the survey may be to produce reliable estimates for each region of a country and the regions may vary in population. To accomplish this goal, it may be necessary to allocate sample sizes to the smaller regions that are substantially greater than would be allocated under proportional stratified sampling. Data-collection costs that differ greatly by strata may offer another reason for deviating from a proportional allocation. An optimal design in this case would be one that allocates larger-than-proportional sample sizes to the strata with lower data-collection costs.

15. The gain in precision derived from proportionate stratification does not necessarily apply with respect to a disproportionate allocation of the sample. To simplify the discussion for this case, we assume that the within-stratum population variances are constant, in other words, that $S_h^2 = S_c^2$ for all strata. This assumption is often a reasonable one in national household surveys when disproportionate stratification is used for the reasons given above. Under this assumption, equation (10) simplifies to

$$V\left(\bar{y}_{st}\right) = S_c^2 \Sigma_h \frac{W_h^2}{n_h} = \frac{S_c^2}{N} \Sigma_h W_h w_h \tag{13}$$

The design effect in this case is

$$D^2(\bar{y}_{st}) = \frac{S_c^2}{S^2} \frac{n}{N} \Sigma_h W_h w_h \tag{14}$$

16. In addition to assuming constant within-stratum variances as used in deriving equation (14), it is often reasonable to assume that stratum means are approximately equal, that is to say, that $\bar{Y}_h = \bar{Y}$ for all strata. With this further assumption, $S_c^2 = S^2$ and the design effect reduces to

$$D^2(\bar{y}_{st}) = \frac{n}{N}\sum_h W_h w_h = n\sum_h \frac{W_h^2}{n_h} \tag{15}$$

Kish (1992)[19] presents the design effect due to disproportionate allocation as

$$D^2(\bar{y}_{st}) = (\sum_h W_h w_h)(\sum_h W_h / w_h) \tag{16}$$

This formula is a very useful one for sample design. However, it should not be applied uncritically without attention to the reasonableness of its underlying assumptions (see below).

17. For a simple example of the application of equation (16), consider a country with two regions where the first region contains 80 per cent of the total population and the second region contains 20 per cent (hence $W_1 = 4W_2$). Suppose that a survey is conducted with equal sample sizes allocated to the two regions ($n_1 = n_2 = 1,000$). Any of the above expressions can be used to compute the design effect from the disproportionate allocation for the estimated national mean (assuming that the means and unit variances are the same in the two regions). For example, using equation (16) and noting that $w_1 = 4w_2$, the design effect is

$$D_w^2(\bar{y}_{st}) = (4W_2 \cdot 4w_2 + W_2 \cdot w_2)\left(\frac{4W_2}{4w_2} + \frac{W_2}{w_2}\right) = 1.36$$

since $W_2 = 0.2$. The disproportionate allocation used to achieve approximately equal precision for estimates from each of the regions results in an estimated mean for the entire country with an effective sample size of $n_{eff} = 2,000/1.36 = 1,471.$

18. Table VI.1 shows the design effect due to disproportionate allocation for some commonly used over-sampling rates when there are only two strata. The figures at the head of each column are the ratios of the weights in the two strata, which are equivalent to inverses of the ratios of the sampling rates in the two strata. The stub items are the proportions of the population in the first stratum. Since the design effect is symmetric around 0.50, values for $W_1 > 0.5$ can be obtained by using the row corresponding to $(1 - W_1)$. To illustrate the use of the table, consider the example given above. The value in the row where W_1 = 0.20 and the column where the over-sampling ratio is 4 gives $D^2(\bar{y}_{st}) = 1.36$. The table shows that the design effects increase as the ratio of the sampling rates increase and the proportion of the population in the strata approaches 50 per cent. When the sampling rates in the strata are very different, then the design effect for the overall mean can be very large and hence the effective sample size is small. The disproportionate allocation results in a very inefficient sample for estimating the overall population statistic in this case.

[19] This reference summarizes many of the results in very useful form. Many of the relationships had been well known and were published decades earlier. See, for example, Kish (1965) and Kish (1976).

19. Many national surveys are intended to produce national estimates and also estimates for various regions of the country. Usually, the regions vary markedly in size. In this situation, a conflict arises in determining an appropriate sample allocation across the regions, as indicated by the above results. Under the assumptions of equal means and unit variances within regions, the optimal allocation for national estimates is a proportionate allocation, whereas for regional estimates it is an equal sample size in each region. The use of the optimal allocation for one purpose will result in a poor sample for the other. A compromise allocation may, however, work reasonably well for both purposes (see sect. D).

Table VI.1. Design effects due to disproportionate sampling in the two-strata case

	Ratio of w_1 to w_2							
W_1	1	2	3	4	5	8	10	20
0.05	1.00	1.02	1.06	1.11	1.15	1.29	1.38	1.86
0.10	1.00	1.05	1.12	1.20	1.29	1.55	1.73	2.62
0.15	1.00	1.06	1.17	1.29	1.41	1.78	2.03	3.30
0.20	1.00	1.08	1.21	1.36	1.51	1.98	2.30	3.89
0.25	1.00	1.09	1.25	1.42	1.60	2.15	2.52	4.38
0.35	1.00	1.11	1.30	1.51	1.73	2.39	2.84	5.11
0.50	1.00	1.13	1.33	1.56	1.80	2.53	3.03	5.51

20. Equation (16) is widely used in sample design to assess the effect of the use of a disproportionate allocation on national estimates. In employing it, however, users should pay attention to the assumptions of equal within-stratum means and variances on which it is based. Consider first the situation where the means are different but the variances are not. In this case, the design effect from disproportionate stratification is given by equation (14), with the additional factor S_c^2 / S^2. This factor is less that 1, and hence the design effect is not as large as that given by equation (16). The design effect, however, represents the overall effect of the stratification and the disproportionate allocation. To measure just the effect of the disproportionate allocation, the appropriate comparison is between the disproportionate stratified sample and a proportionate stratified sample of the same size. The ratio of the variance of $\bar{y}_{st}$ for the disproportionate design to that of $\bar{y}_{prop}$ is, from equations (11) and (13) with $S_w^2 = S_c^2$,

$$R = V(\bar{y}_{st}) / V(\bar{y}_{prop}) = (\Sigma_h W_h w_h)(\Sigma_h W_h / w_h)$$

Thus, in this case, the formula in equation (16) can be interpreted as the effect of just the disproportionate allocation.

21. The assumption of equal within-stratum unit variances is more critical. The above results show that a disproportionate allocation leads to a loss of precision in overall estimates when within-stratum unit variances are equal, but this does not necessarily hold when the within-

stratum unit variances are unequal. Indeed, when within-stratum variances are unequal, the optimum sampling fractions to be used are proportional to the standard deviations in the strata [see, for example, Cochran (1977)]. This type of disproportionate allocation is widely used in business surveys. It can lead to substantial gains in precision over a proportionate allocation when the within-stratum standard deviations differ markedly.

22. In household surveys, the assumption of equal, or approximately equal, within-stratum variances is often reasonable. One type of estimate for which the within-stratum variances may be unequal is a proportion. A proportion is the mean of a variable that takes on only the values 1 and 0, corresponding to having or not having the given characteristic. The unit variance for such a variable is $\sigma^2 = P(1-P)$, where P is the population proportion with the characteristic. Thus, the unit variance in stratum *h* with a proportion P_h having the characteristic is $S_h^2 = P_h(1-P_h)$. If P_h varies across strata, so will S_h^2. However, the variation in S_h^2 is only slight for proportions between 0.2 and 0.8, from a high of 0.25 for $P_h = 0.5$ to a low of 0.16 for $P_h = 0.2$ or 0.8.

23. To illustrate the effect of variability in stratum proportions and hence in stratum variances, we return to our example with two strata with $W_1 = 0.8$, $W_2 = 0.2$ and $n_1 = n_2$, and consider two different sets of values for P_1 and P_2. For case 1, let $P_1 = 0.5$ and $P_2 = 0.8$. Then the overall design effect, computed using equations (10) and (1), is $D^2(\bar{y}_{st}) = 1.35$ and the ratio of the variances for the disproportionate and proportionate designs is $R = 1.43$. For case 2, let $P_1 = 0.8$ and $P_2 = 0.5$. Then $D^2(\bar{y}_{st}) = 1.16$ and $R = 1.26$. The values obtained for $D^2(\bar{y}_{st})$ and *R* in these two cases can be compared with the design effect of 1.36 that was obtained under the assumption of equal within-stratum variances. In both cases, the overall design effects are less than 1.36 because of the gain in precision from the stratification. In case 1, the value of *R* is greater than 1.36, because stratum 1, which is sampled at the lower rate, has the larger within-stratum variance. In case 2, the reverse holds: stratum 2, which is over-sampled, has the larger within-stratum variance. This oversampling is therefore in the direction called for to give increased precision. In fact, in this case the optimal allocation would be to sample stratum 2 at a rate 1.25 times as large as the rate in stratum 1. Even though the stratum proportions differ greatly in these examples and, as a consequence, the within-stratum variances also differ appreciably, the values of *R* obtained – at 1.26 and 1.43 – are reasonably close to 1.36. These calculations illustrate the fact that the approximate measure of the design effect from weighting produced from equation (16) is adequate for most planning purposes even when the within-stratum variances differ to some degree.

24. Finally, consider a more extreme example with $P_1 = 0.05$ and $P_2 = 0.5$, still with $W_1 = 0.8$, $W_2 = 0.2$ and $n_1 = n_2$. In this case, $D^2(\bar{y}_{st}) = 0.67$ and $R = 0.92$. This example demonstrates that disproportionate stratification can produce gains in precision. However, given the assumptions on which it is based, equation (16) cannot produce a value less than 1. Thus, equation (16) should not be applied indiscriminately without attention to its underlying assumptions.

2. Clustering

25. We now consider another major component of the overall design effect in most general population surveys, namely, the design effect due to clustering in multistage samples. Samples are clustered to reduce data-collection costs since it is uneconomical to list and sample households spread thinly across an entire country or region. Typically, two or more stages of sampling are employed, where the first-stage or primary sampling units (PSUs) are clearly defined geographical areas that are generally sampled with probabilities proportional to the estimated numbers of households or persons that they contain. Within the selected PSUs, one or more additional stages of area sampling may be conducted and then, in the sub-areas finally selected, dwelling units are listed and households are sampled from the lists. For a survey of households, data are collected for sampled households. For a survey of persons, a list of persons is compiled for selected households and either all or a sample of persons eligible for the survey is selected. For the purposes of this discussion, we assume a household survey with only two stages of sampling (PSUs and households). However, the extension to multiple stages is direct.

26. In practical settings, PSUs are always variable in size (that is to say, in the numbers of units they contain) and for this reason they are sampled by probability proportional to estimated size (PPES) sampling. The sample sizes selected from selected PSUs also generally vary between PSUs. However, for simplicity, we start by assuming that the population consists of A PSUs (for example, census enumeration districts) each of which contains B households. A simple random sample of a PSUs is selected and a simple random sample of $b \le B$ households is selected in each selected PSU (the special case when $b = B$ represents a single-stage cluster sample). We assume that the first-stage finite population correction factor is negligible. The sample design for selecting households uses the equal probability of selection method (epsem), so that the population mean can be estimated by the simple unweighted sample mean $\bar{y}_{cl} = \sum_{\alpha}^{a} \sum_{\beta}^{b} y_{\alpha\beta} / n$, where $n = ab$ and the subscript cl denotes the cluster. The variance of $\bar{y}_{cl}$ can be written as

$$V\left(\bar{y}_{cl}\right) = \frac{S^2}{n}\left[1 + (b-1)\,\rho\right] \tag{17}$$

where S^2 is the unit variance in the population and ρ is the intra-class correlation coefficient that measures the homogeneity of the y-variable in the PSUs. In practice, units within a PSU tend to be somewhat similar to each other for nearly all variables, although the degree of similarity is usually low. Hence, ρ is almost always positive and small.

27. The design effect in this simple situation is

$$D^2\left(\bar{y}_{cl}\right) = 1 + (b-1)\rho \tag{18}$$

This basic result shows that the design effect from clustering the sample within PSUs depends on two factors: the subsample size within selected PSUs (b) and the intra-class correlation (ρ). Since ρ is generally positive, the design effect from clustering is, as a rule, greater than 1.

28. An important feature of equation (18) - and others like it presented below - is that it depends on ρ which is a measure of homogeneity within PSUs for a particular variable.[20] The value of ρ is near zero for many variables (for example, age and sex), and small but non-negligible for others (for example, $\rho = 0.03$ to 0.05), but it can be high for some (for example, access to a clinic in the village - the PSU - when all persons in a village will either have or not have access). It is theoretically possible for ρ to be negative, but this is unlikely to be encountered in practice (although sample estimates of ρ are often negative). Frequently, ρ is inversely related to the size of the PSU because larger clusters tend to be more diverse, especially when PSUs are geographical areas. These types of relationships are exploited in the optimal design of surveys, where PSUs that are large and more diverse are used when there is an option. Estimates of ρ for key survey variables are needed for planning sample designs. These estimates are usually based on estimates from previous surveys for the same or similar variables and PSUs, and the belief in the portability of the values of ρ across similar variables and PSUs.

29. In real settings, PSUs are not of equal size and they are not sampled by simple random sampling. In most national household sample designs, stratified samples of PSUs are selected using PPES sampling. As a result, equation (18) does not directly apply. However, it still serves as a useful model for the design effect from clustering for a variety of epsem sample designs with a suitable modification with respect to the interpretation of ρ.

30. Consider first an unstratified PPS sample of PSUs, where the exact measures of size are known. In this case, the combination of a PPS sample of *a* PSUs and an epsem sample of *b* households from each sampled PSU produces an overall epsem design. With such a design, equation (18) still holds, but with ρ now interpreted as a synthetic measure of homogeneity within the ultimate clusters created by the subsample design (Kalton, 1979). The value of ρ, for instance, for a subsample design that selects *b* households by systematic sampling is different from that for a subsample design that divides each sampled PSU into sub-areas containing *b* households each and selects one sub-area (the value of ρ is likely to be larger in the latter case). This extension thus deals with both PPS sampling and with various alternative forms of subsample design.

31. Now consider stratification of the PSUs. Kalton (1979) shows that the design effect due to clustering in an overall epsem design in which a stratified sample of *a* PSUs is selected and *b* elementary units are sampled with equal probability within each of the selected PSUs can be approximated by

$$D^2(\bar{y}_{cl}) = 1 + (b-1)\bar{\rho} \tag{19}$$

where $\bar{\rho}$ is the average within-stratum measure of homogeneity, provided that the homogeneity within each stratum is roughly of the same magnitude. The gain from effective stratification of PSUs can be substantial when *b* is sizeable because the overall measure of homogeneity in (18) is replaced by a smaller within-stratum measure of homogeneity in equation (19). Expressed

[20] The discussion in the present section applies to the measure of within-cluster homogeneity for both equal- and unequal-sized clusters.

otherwise, the reduction in the design effect of $(b-1)(\rho-\bar{\rho})$ from stratified sampling of the PSUs can be large when *b* is sizeable.

32. Thus far, we have assumed an overall epsem sample in which the sample size in each selected PSU is the same, *b*. These conditions are met when equal-sized PSUs are sampled with equal probability and when unequal-sized PSUs are sampled by exact PPS sampling. However, in practice neither of these situations applies. Rather unequal-sized PSUs are sampled by PPES, with estimated measures of size that are inaccurate to some degree. In this case, the application of the subsampling rates in the sampled PSUs to give an overall epsem design results in some variation in subsample size. Provided that the variation in the subsample sizes is not large, equation (19) may still be used as an approximation, with *b* being replaced by the average subsample size, that is to say,

$$D^2(\bar{y}_{cl}) = 1 + (\bar{b} - 1)\bar{\rho} \tag{20}$$

where $\bar{b} = \sum b_\alpha / a$ and b_α is the number of elementary units in PSU α. Equation (20) has proved to be of great practical utility for situations in which the number of sampled units in each of the PSUs is relatively constant.

33. When the variation in the subsample sizes per PSU is substantial, however, the approximation involved in equation (20) becomes inadequate. Holt (1980) extends the above approximation to deal with unequal subsample sizes by replacing $\bar{b}$ in equation (20) by a weighted average subsample size. The design effect due to clustering with unequal cluster sizes can be written as

$$D^2(\bar{y}_{cl}) = 1 + (b' - 1)\bar{\rho} \tag{21}$$

where $b' = \sum b_\alpha^2 \big/ \sum b_\alpha$. (The quantity b' can be thought of as the weighted average $b' = \Sigma k_\alpha b_\alpha / \Sigma k_\alpha$, where $k_\alpha = b_\alpha$.) As above, the approximation assumes an overall epsem sample design.

34. As an example, suppose that there are five sampled PSUs with subsample sizes of 10, 10, 20, 20 and 40 households, and suppose that $\bar{\rho} = 0.05$. The average subsample size is $\bar{b} = 20$, whereas $b' = 26$. In this example, the design effect due to clustering is thus 1.95 using approximation (20) as compared with 2.25 using approximation (21).

35. Verma, Scott and O'Muircheartaigh (1980) and Verma and Lê (1996) provide another way of writing this adjustment that is appropriate when subsample sizes are very different for different domains (for example, urban and rural domains). With two domains, suppose that b_1 households are sampled in each of a_1 sampled PSUs in one domain, with $n_1 = a_1 b_1$, and that b_2 households are sampled in the remaining a_2 sampled PSUs in the other domain, with $n_2 = a_2 b_2$. Then, with this notation,

$$b' = (n_1 b_1 + n_2 b_2)/(n_1 + n_2)$$

36. The preceding discussion has considered the design effects from clustering for estimates of means (and proportions) for the total population. Much of the treatment is equally applicable to subgroup estimates, provided that there is careful attention to the underlying assumptions. It is useful to introduce a threefold classification of types of subgroups according to their distributions across the PSUs. At one end, there are subgroups that are evenly spread across the PSUs that are known as "cross-classes." For example, age/sex subgroups are generally cross-classes. At the other end, there are subgroups, each of which is concentrated in a subset of PSUs, that are termed "segregated classes." Urban and rural subgroups are likely to be of this type. In between are subgroups that are somewhat concentrated by PSU. These are "mixed classes".

37. Cross-classes follow the distribution of the total sample across the PSUs. If the total sample is fairly evenly distributed across the PSUs, then equation (20) may be used to compute an approximate design effect from clustering and that equation may also be used for a cross-class. However, when it is applied for a cross-class, an important change arises: $\overline{b}$ now represents the average cross-class subsample size per PSU. As a result of this change, design effects for cross-class estimates are smaller than those for total sample estimates.

38. Segregated classes constitute all the units in a subset of the PSUs in the full sample. Since the subclass sample size for a segregated class is the same as that for the total sample in that subset of PSUs, in general, there is no reason to expect the design effect for an estimate for a segregated class to be lower than that for a total sample estimate. The design effect for an estimate for a segregated class will differ from that for a total sample estimate only if the average subsample size per PSU in the segregated class differs from that in the total sample or if the homogeneity differs (including, for example, a difference in the synthetic ρ due to different subsample designs in the segregated class and elsewhere). If the total sample is evenly spread across the PSUs, equation (20) may again be applied, with $\overline{b}$ and ρ being values for the set of PSUs in the segregated class.

39. The uneven distribution of a mixed class across the PSUs implies that equation (20) is not applicable in this case. For estimating the design effect from clustering for an estimate from a mixed class, equation (21) may be used, with b_α being the number of sampled members of the mixed class in PSU α.

3. Weighting adjustments

40. As discussed in section B.1, entitled "Stratification", the unequal selection probabilities between strata with disproportionate stratification result in a need to use weights in the analysis of the survey data. Equations (15) and (16) give the design effect arising from the disproportionate stratification and resulting unequal weights under the assumptions that the strata means and unit variances are all equal. We now turn to alternative forms of these formulae that are more readily applied to determine the effects of weights at the analysis stage. First, however, we note the factors that give rise to the need for variable weights in survey analysis [see also Kish (1992)]. In the first place, as we have already noted, variable weights are needed in the

analysis to compensate for unequal selection probabilities associated with disproportionate stratification. More generally, they are needed to compensate for unequal selection probabilities arising from any cause. The weights that compensate for unequal selection probabilities are the inverses of the selection probabilities, and they are often known as base weights. The base weights are often then adjusted to compensate for non-response and to make weighted sample totals conform to known population totals. As a result, final analysis weights are almost always variable to some degree.

41. Even without oversampling of certain domains, sample designs usually deviate from epsem because of frame problems. For example, if households are selected with equal probability from a frame of households and then one household member is selected at random in each selected household, household members are sampled with unequal probabilities and hence weights are needed in the analysis in compensation. These weights give rise to a design effect component as discussed below. In passing, it may be noted that this weighting effect may be avoided by taking all members of selected household into the sample. However, this procedure introduces another stage of clustering, with an added clustering effect due to the similarity of many characteristics of household members [see Clark and Steel (2002) on the design effects associated with these alternative methods of selecting persons in sampled households].

42. Another common case of a non-epsem design resulting from a frame problem is that in which a two-stage sample design is used and the primary sampling units (PSUs) are sampled with probabilities proportional to estimated sizes (PPES). If the size measures are reasonably accurate, the sample size per selected PSU for an overall epsem design is roughly the same for all PSUs. However, if the estimated size of a selected PSU is a serious underestimate, the epsem design calls for a much larger than average number of units from that PSU. Since collecting survey data for such a large number is often not feasible, a smaller sample may be drawn, leading to unequal selection probabilities and the need for compensatory weights.

43. Virtually all surveys encounter some amount of non-response. A common approach used to reduce possible non-response bias involves differentially adjusting the base weights of the respondents. The procedure consists of identifying subgroups of the sample that have different response rates and inflating the weights of respondents in each subgroup by the inverse of the response rate in that subgroup (Brick and Kalton, 1996). These weighting adjustments cause the weights to vary from the base weights and the effect is often an increase in the design effect of an estimate.

44. When related population information is available from some other source, the non-response-adjusted weights may be further adjusted to make the weighted sample estimates conform to the population information. For example, if good estimates of regional population sizes are available from an external source, the sample estimates of these regional populations can be made to coincide with the external estimates. This kind of population weighting adjustment is often made by a post-stratification type of adjustment. It can help to compensate for non-coverage and can improve the precision of some survey estimates. However, it adds further variability to the weights which can adversely affect the precision of survey estimates that are unrelated to the population variables employed in the adjustment.

45. With this background, we now consider a generalization of the design effect for disproportionate stratification to assess the general effects of variable weights. Kish (1992) presents another way of expressing the design effect for a stratified mean that is very useful for computing the effect of disproportionate stratification at the analysis stage. The following equation is simply a different representation of equations (15) and (16), and is thus based on the same assumptions of equal strata means and unit variances, particularly the latter. Since it is computed from the sample, the design effect is designated as $d^2(\bar{y}_{st})$ and

$$d^2(\bar{y}_{st}) = \frac{n\sum_h \sum_i w_{hi}^2}{\left(\sum_h \sum_i w_{hi}\right)^2} = 1 + cv^2\left(w_{hi}\right) \tag{22}$$

where $cv(w_{hi})$ is the coefficient of variation of the weights, $cv^2\left(w_{hi}\right) = \sum\sum\left(w_{hi} - \bar{w}\right)^2 / n\bar{w}^2$, and $\bar{w} = \sum\sum w_{hi} / n$ is the mean of the weights.

46. A more general form of this equation is given by

$$d^2(\bar{y}_{st}) = \frac{n\sum_j w_j^2}{\left(\sum_j w_j\right)^2} = 1 + cv^2(w_j) \tag{23}$$

where each of the n units in the sample has its own weight w_j ($j = 1, 2, \ldots, n$). The design effect due to unequal weighting given by equation (23) depends on the assumption that the weights are unrelated to the survey variable. The equation can provide a reasonable measure of the effect of differential weighting for unequal selection probabilities if its underlying assumptions hold at least approximately [see Spencer (2000), for an approximate design effect for the case where the selection probabilities are correlated with the survey variable].

47. Non-response adjustments are generally made within classes defined by auxiliary variables known for both respondents and non-respondents. To be effective in reducing non-response bias, the variables measured in the survey do need to vary across these weighting classes. The variation, however, is generally not great, particularly in the unit variance. As a result, equation (23) is widely used to examine the effect of non-response weighting adjustments on the precision of survey estimates. This examination may be conducted by computing equation (23) with the base weights alone or with the non-response adjustment weights. If the latter computation produces a much larger value than the former, this means that the non-response weighting adjustments are causing a substantial loss of precision in the survey estimates. In this case, it may be advisable to modify the weighting adjustments by collapsing weighting classes or trimming extremely large weights in order to reduce the loss of precision.

48. While equation (23) is reasonable with respect to most non-response sample weighting adjustments, it often does not yield a good approximation for the effect of population weighting adjustments. In particular, when the weights are post-stratified or calibrated to known control totals from an external source, then the design effect for the mean of y is poorly approximated by

equation (23) when y is highly correlated with the one or more of the control totals. For example, assume the weights are post-stratified to control totals of the numbers of persons in a country by sex. Consider the extreme case where the survey data are used to estimate the proportion of women in the population. In this case of perfect correlation between the y variable and the control variable, the estimated proportion is not subject to sampling error and hence has zero variance. In practice, the correlation will not be perfect, but it may be sizeable for some of the survey variables. When the correlation is sizeable, post-stratification or calibration to known population totals can appreciably improve the precision of the survey estimates, but this improvement will not be shown through the use of equation (23). On the contrary, equation (23) will indicate a loss in precision.

49. The above discussion indicates that equation (23) should not be used to estimate the design effects from population weighting adjustments for estimates based on variables that are closely related to the control variables. In most general population surveys in developing countries, however, few, if any, dependable control variables are available, and the relationships between any that are available and the survey variables are seldom strong. As a result, the problem of substantially overestimating the design effects from weighting using equation (23) should not occur often. Nevertheless, the above discussion provides a warning that equation (23) should not be applied uncritically.

50. We conclude this discussion of the design effects of weighting with some comments on the effects of weighting on subgroup estimates. All the results presented in this section and section B.1 can be applied straightforwardly to give the design effects for subgroup estimates simply by restricting the calculations to subgroup members. However, care must be taken in trying to infer the design effects from weighting for subgroup estimates from results for the full sample. For this inference to be valid, the distribution of weights in the subgroup must be similar to that in the full sample. Sometimes this is the case, but not always. In particular, when disproportionate stratification is used to give adequate sample sizes for certain domains (subgroups), the design effects for total sample estimates will exceed 1 (under the assumptions of equal means and variances). However, the design effects from weighting for domain estimates may equal 1 because equal selection probabilities are used within domains.

C. Models for design effects

51. The previous section has presented some results for design effects associated with weighting and clustering separately, with the primary focus on design effects for means and proportions. The present section extends those results by considering the design effects from a combination of weighting and clustering and the design effects for some other types of estimates.

52. A number of models have been used to represent the design effects for these extensions. The models have been used in both the design and the analysis of complex sample designs (Kalton, 1977; Wolter, 1985). Historically, the models have played a major role in analysis. However, their use in analysis is probably on the wane. Their primary -- and important -- use in the future, in the planning of new designs, will be the focus of the present discussion.

53. Recent years have seen major advances in computing power and in software for computing sampling errors from complex sample designs. Before these advances were achieved, computing valid sampling errors for estimates from complex samples had been a laborious and time-consuming task. It was therefore common practice to compute sampling errors directly for only a relatively small number of estimates and to use design effect or other models to infer the sampling errors for other estimates. The computing situation has now improved dramatically so that the direct computation of sampling errors for many estimates is no longer a major hurdle. Moreover, further improvements in both computing power and software can be expected in the future. Thus, the use of design effects models for this purpose can be expected to largely disappear.

54. Another reason for using sampling error models at the analysis stage is to provide a means for succinctly summarizing sampling errors in survey reports, thereby eliminating the need to present a sampling error for each individual estimate. In some cases, it may also be argued that the sampling error estimates from a model may be preferable to direct sampling error estimates because they are more precise. There are certain cases where this latter argument has some force (for instance, in estimating the sampling error for an estimate in a region in which the number of sampled PSUs is very small). However, in general, the use of models for reporting sampling errors for either of these reasons is questionable. The validity of the model estimates depends on the validity of the models and, when comparisons of direct and model-based sampling errors have been made, the comparisons have often raised serious doubts about the validity of the models [see, for example, Bye and Gallicchio (1989)]. Also, while sampling error models can provide a concise means of summarizing sampling errors in survey reports, they impose on users the undesirable burden of performing calculations of sampling errors from the models. Our overall conclusion is that design effect and other sampling error models will play a limited role in survey analysis in the future.

55. In contrast, design effect models will continue to play a very important role in sample design. Understanding the consequences of a disproportionate allocation of the sample and of the effects of clustering on the precision of different types of survey estimates is key to effective sample design. Most obviously, the determination of the sample size required to give adequate precision to key survey estimates clearly needs to take account of the design effect resulting from a given design. Also, the structure of an efficient sample design can be developed by examining the results from models for different designs. Note that estimates of unknown parameters, such as ρ, are required in order to apply the models at the design stage. This requirement points to the need for producing estimates of these parameters from past surveys, as illustrated in the next section.

56. We start by describing models for inferring the effects of clustering in epsem samples on a range of statistics beyond the means and proportions considered in section B.3, entitled "Weighting adjustments". To introduce these models, we return to subgroup means as already discussed, with the distinction made between cross-classes, segregated classes, and mixed classes. For a cross-class, denoted as *d*, that is evenly spread across the PSUs, the design effect for a cross-class mean is given approximately by equation (20), which is written here as

$$D^2(\bar{y}_{cl:d}) = 1 + (\bar{b}_d - 1)\bar{\rho}_d \qquad (24)$$

where $\bar{b}_d$ denotes the average cross-class sample size per PSU and $\bar{\rho}_d$ is the synthetic measure of homogeneity of y in the PSUs for the cross-class. A widely used model assumes that the measure of homogeneity for the cross-class is the same as that for the total population, in other words, that $\bar{\rho}_d = \bar{\rho}$. Then the design effect for the cross-class mean can be estimated by

$$d^2(\bar{y}_{cl:d}) = 1 + (\bar{b}_d - 1)\hat{\rho} \tag{25}$$

where $\hat{\rho}$ is an estimate of $\bar{\rho}$ from the full sample given by

$$\hat{\rho} = \frac{d^2(\bar{y}_{cl}) - 1}{\bar{b} - 1} \tag{26}$$

57. A common extension of this approach is to compute $\hat{\rho}$'s for a set of comparable estimates involving related variables and, provided that the $\hat{\rho}$'s are fairly similar, to use some form of average of them to estimate $\bar{\rho}$ and hence also the $\bar{\rho}_d$'s for subgroup estimates for all the variables. This approach has often been applied to provide design effect models for summarizing sampling errors in survey reports. It is also the basis of one form of generalized variance function (GVF) used for this purpose (Wolter, 1985, p. 204).

58. A special case of this approach occurs with survey estimates that are subgroup proportions falling in different categories of a categorical variable, such as the proportions of different subgroups that have reached different levels of education or that are in different occupational categories. It is often assumed that the values of $\bar{\rho}$ for the different categorizations are similar, so that the value of $\bar{\rho}$ needs to be estimated for only one categorization, and that once estimated, $\hat{\rho}$ can then be applied for all the other categorizations. The assumption of a common $\bar{\rho}$ is mathematically correct when there are only two categories (for example, household with and household without electricity), but it need not hold when there are more than two categories. Consider, for example, estimates of the proportion of workers engaged in agriculture and in mining. The value of $\bar{\rho}$ for agricultural workers is almost certainly much lower than that for miners because mining is probably concentrated in a few areas. The assumption of a common $\bar{\rho}$ value for all categorizations should therefore not be applied uncritically.

59. When variances for cross-class means derived from equation (25) have been compared with those computed directly, they have been found to tend to be underestimates. This finding may be due to the fact that, even though classified as cross-classes, the subgroups are not distributed completely evenly across the PSUs. One remedy that has been used to address this problem is to modify equation (25) with the result that

$$d^2(\bar{y}_{cl:d}) = 1 + k_d(\bar{b}_d - 1)\hat{\rho} \tag{27}$$

where $k_d > 1$. Basing his work on many empirical analyses, Kish (1995) suggests values of $k_d = 1.2$ or 1.3; Verma and Lê (1996) allow k_d to vary with the cross-class size (with k_d always greater than 1). A possible alternative remedy would be to replace $\bar{b}_d$ in (25) with $b'_d = \Sigma b_{d\alpha}^2 / \Sigma b_{d\alpha}$ in line with equation (21).

60. We now consider briefly design effects for analytic statistics. The simplest and most widely used form of analytic statistic is the difference between two subgroup means or proportions. It has generally been found that the design effect for the difference between two means is greater than 1 but less than that obtained by treating the two subgroup means as independent (Kish and Frankel, 1974; Kish, 1995). Expressed in terms of variances,

$$V(\bar{y}_{u:d}) + V(\bar{y}_{u:d'}) < V(\bar{y}_{cl:d} - \bar{y}_{cl:d'}) < V(\bar{y}_{cl:d}) + V(\bar{y}_{cl:d'}) \tag{28}$$

where d and d' represent the two subgroups. The variance of the difference in the means is typically lower than the upper bound when the subgroups are both represented in the same PSUs. This feature results in a covariance between the two means that is virtually always positive, and that positive covariance then reduces the variance of the difference. This effect does not occur when the subgroups are segregated classes that are in different sets of PSUs: in this case, the upper bound applies. Under the assumption that the unit variances in the two subgroups are the same (in other words, that $S_d^2 = S_{d'}^2$), this inequality reduces to

$$1 < D^2(\bar{y}_d - \bar{y}_{d'}) < \frac{n_{d'} D^2(\bar{y}_d) + n_d D^2(\bar{y}_{d'})}{n_d + n_{d'}}$$

61. A special case of the difference between two proportions arises when the proportions are each based on the same multi-category variable, as occurs, for example, when respondents are asked to make a choice between several alternatives and the analyst is interested in whether one alternative is more popular than another. Kish and others (1995) examined design effects for such differences and found empirically that $d^2(p_{d-}\rho_{d'}) = \left[d^2(\rho_d) + d^2(\rho_{d'})\right]/4$ in this special case.

62. The finding given above that design effects from clustering are typically smaller for differences in means than for overall means generalizes to other analytic statistics. See Kish and Frankel (1974) for some early empirical evidence and some modelling suggestions for design effects for multiple regression coefficients. The design effects for regression coefficients are like those for differences between means. That this is in line with expectation may be seen by noting that the slope of a simple linear regression of y on x may be estimated fairly efficiently by $b = (\bar{y}_u - \bar{y}_l)/(\bar{x}_u - x_l)$, where the means of y and x are computed for the upper (u) and lower (l) thirds of the sample based on the x variable. See Skinner, Holt and Smith (1989) and Lehtonen and Pahkinen (1994) for design effects in regression and other forms of analysis, and Korn and Graubard (1999) for the effects of complex sample designs on precision in the analysis of survey data.

63. We conclude this section with some comments on the taxing problem of decomposing an overall design effect into components due to weighting and to clustering. The calculation of the design effect $d^2(\bar{y}) = v_c(\bar{y})/v_u(\bar{y})$ encompasses the combined effects of weighting and clustering. However, in using the data from the current survey to plan a future survey, the two components of the design effect need to be separated. For example, the future survey may be planned as one using epsem whereas the current survey may have oversampled certain domains. Also, even if it used the same PSUs and stratification, the future survey might wish to change the subsample size per PSU. Kish (1995) discusses this issue, for which there is no single and simple solution. Here, we give an approach that may be used only when the weights are random or approximately so. In this case, the overall design effect can be decomposed approximately into a product of the design effects of weighting and clustering whereby

$$d^2(\bar{y}) = d_w^2(\bar{y}).d_{cl}^2(\bar{y}) \tag{29}$$

where $d_w^2(\bar{y})$ is the design effect from weighting as given by equation (23) and $d_{cl}^2(\bar{y})$ is the design effect from clustering given by equations (20) or (21). There is little theoretical justification for equation (29); however, using a modelling approach, Gabler, Haeder and Lahiri (1999) derive the design effect given by equation (29) as an upper bound. Using equation (29) with equation (20), $\bar{\rho}$ is thus estimated by

$$\hat{\bar{\rho}} = \frac{[d^2(\bar{y})/d_w^2(\bar{y})]-1}{\bar{b}-1} \tag{30}$$

As will be seen below, for planning purposes, estimation of the parameter $\bar{\rho}$ is more important than estimation of the design effect from clustering because it is more portable across different designs. The design effect from clustering in one survey can be directly applied in planning another only if the subsample size per PSU remains unchanged.

D. Use of design effects in sample design

64. The models for design effects discussed in the earlier part of this chapter can serve as useful tools for planning a new sample design. However, they need to be supported by empirical data, particularly on the synthetic measure of homogeneity $\bar{\rho}$. These data can be obtained by analysing design effects for similar past surveys. Accumulation of data on design effects is therefore valuable.

65. A substantial amount of data on design effects is available for demographic surveys of fertility and health from the extensive analyses of sampling errors that have been conducted for the World Fertility Surveys (WFS) and Demographic and Health Surveys (DHS) programmes. The WFS programme had conducted 42 surveys in 41 countries between 1974 and 1982. The DHS programme followed in 1984, with over 120 completed surveys in 66 countries having been conducted to date, with the surveys being repeated in most countries every three to five years. See Verma and Lê (1996) for analyses of DHS sampling errors, and Kish, Groves and

Krotki (1976) and Verma, Scott and O'Muircheartaigh (1980) for similar analyses of WFS sampling errors. An important finding from the sampling error analyses for these programmes is that estimates of $\bar{\rho}$ for a given estimate are fairly portable across countries provided that the sample designs are comparable. Thus, in designing a new survey in one country, empirical data on sampling errors from a similar survey in a neighbouring country may be employed if necessary and if due care is taken to check on sample design comparability.

66. The example given below illustrates the use of design effects in developing the sample design for a hypothetical national survey. For the purposes of this illustration, we assume that the sample design will be a stratified two-stage PPS sample, say, with census enumeration districts as the PSUs and households as the second-stage units. We assume that the key statistic of interest is the proportion of households in poverty, which for planning purposes is assumed to be about 25 per cent, and to be similar for all the provinces in the country. The initial specifications are that the estimate of this proportion should have a coefficient of variation of no more that 5 per cent for the nation and no more than 10 per cent for each of the nation's eight provinces. Furthermore, the sample should be efficient in producing precise estimates for a range of statistics for national subgroups that are spread fairly evenly across the eight provinces. If simple random sampling was used, the coefficient of variation would be

$$CV = \sqrt{\frac{1-P}{nP}}$$

where P is the proportion of households in poverty (25 per cent in this case). This formula can also be used with a complex sample design, but with n replaced by the effective sample size, $n_{eff} = n / D^2(p)$.

67. The first issue to be addressed is how the sample should be distributed across the provinces. Table VI.2 gives the distribution of the population across the provinces (W_h), together with a proportionate allocation of the sample across the provinces, an equal sample size allocation for each province, and a compromise sample allocation that falls between the proportionate and equal allocations. An arbitrary total sample size of 5,000 households is used at this point. It can be revised later, if necessary.

Table VI.2. Distributions of the population and three alternative sample allocations across the eight provinces (A –H)

	A	B	C	D	E	F	G	H	Total
W_h	0.33	0.24	0.20	0.10	0.05	0.04	0.02	0.02	1.00
Proportionate allocation	1 650	1 200	1 000	500	250	200	100	100	5 000
Equal sample size allocation	625	625	625	625	625	625	625	625	5 000
Compromise sample allocation	1 147	879	767	520	438	427	411	411	5 000

68. Other things being equal, the proportionate allocation is the most suitable for producing national estimates and subgroup estimates where the subgroups are evenly spread across the provinces. On the other hand, the equal sample size allocation is the most suitable for producing provincial estimates. As table VI.2 shows, these two allocations differ markedly, as a result of the very different sizes of the provinces given in the W_h row. The proportionate allocation yields samples in the small provinces (E, F, G and H) that are too small to enable the computation of reliable estimates for them. On the other hand, the equal sample size allocation reduces the precision of national estimates. That loss of precision can be computed from equation (15), which, in this case, simplifies to $H\Sigma W_h^2 = 1.77$, where H is the number of provinces. Thus, by considering the effects of the disproportionate allocation only (that is to say, by excluding the effects of clustering), the sample size of 5,000 for national estimates is reduced to an effective sample size of $5,000/1.77 = 2,825$.

69. Whether the large loss of precision for national estimates (particularly for subgroups) resulting from the use of the equal allocation is acceptable depends on the relative importance of national and provincial estimates. Often, national estimates are sufficiently important to render this loss too great to accept. In this case, a compromise allocation that falls between the proportionate and equal allocations may be found to satisfy the needs for both national and provincial estimates. The compromise allocation in the final row of table VI.2 is computed according to an allocation proposed by Kish (1976, 1988) for the situation where national and provincial estimates are of equal importance. That allocation, given by $n_h \propto \sqrt{W_h^2 + H^{-2}}$, increases the sample sizes for the small provinces considerably over the proportionate allocation, but not as much as the equal allocation. The design effect for unequal weighting for this allocation is 1.22, as compared with 1.77 for the equal sample size allocation. We will assume that the compromise allocation is adopted for the survey.

70. The next issue to be addressed is how to determine the number of PSUs and the desired number of households to be selected per PSU. As discussed in chapter II, through the use of a simple cost model, the optimum number of households to select per sampled PSU is given by

$$b_{opt} = \sqrt{C^* \frac{(1-\rho)}{\rho}}$$

where C* is the ratio of the cost of adding a PSU to the sample to the cost of adding a household. The cost model is oversimplified, and the formula for b_{opt} should not be used uncritically; nevertheless, it can still give useful guidance.

71. Let us assume that the organizational structure of the survey fieldwork makes the use of the simple cost model reasonable and that an analysis of the cost structure indicates that C^* is about 16. Furthermore, let us assume that a previous survey, using the same PSUs, has produced an estimate of $\bar{\rho} = 0.05$ for a characteristic that is highly correlated with poverty. Applying these numbers to the above formula gives $\hat{b}_{opt} = 17.4$, which, for the sake of simplicity, we round to 17. Often, in practice, the cost ratio C^* is not constant across the country; for

example, the ratio may be much lower in urban than in rural areas. If this is the case, different values may be used in different parts of the country. Such complexity will not be considered further here. Examples of such differences are to be found in several of the chapters in this publication that describe national sample designs.

72. With $\bar{\rho} = 0.05$ and $b = 17$, the design effect from clustering is

$$D^2(p) = 1 + (b-1)\bar{\rho} = 1.80$$

This design effect needs to be taken into account in determining the precision of provincial estimates. For example, the effective sample size of 411 households in province H is $411/1.80 = 228$. Hence, the coefficient of variation for the proportion of households in poverty in province H is 0.11. If this level of precision was deemed inadequate, the sample size in province H (and also G) would need to be increased.

73. The design effect for national estimates needs to combine the design effects for clustering and the disproportionate allocation across provinces. Thus, for the overall national proportion of households in poverty, the estimated design effect may be obtained from equation (29) as $1.22 \times 1.80 = 2.20$. Hence, the effective sample size corresponding to an actual sample size of 5,000 households is 2,277 and the coefficient of variation for the national estimate of the proportion of households in poverty is 0.036. It is often the case that the overall sample size is more than adequate to satisfy the precision requirements for estimates for the total population. Of more concern is the precision levels for population subgroups. In this case, the design effect from clustering for cross-classes evenly distributed across the PSUs, is smaller than for the total sample, as described in section C. For example consider a cross-class that comprises one third of the population. In this case, applying formula (27) with $k_d = 1.2$ and $\bar{b}_d = 17/3$ gives a clustering design effect of 1.23. Combining the clustering design effect with that for the disproportionate allocation across provinces gives an overall design effect for the cross-class estimate of $1.22 \times 1.23 = 1.50$, and an effective sample size of $5000/(3 \times 1.50) = 1111$. The estimated coefficient of variation for the cross-class estimate is thus 0.05.

74. Calculations along the lines of those indicated above can be made to assess the likely precision of key survey estimates, and sample sizes can be modified to meet desired requirements. In the final estimates of sample sizes, allowances need to be made for non-response. For example, with a fairly uniform 90 per cent response rate across the country, the sample sizes calculated above need to be increased by 11 per cent. Also, the design effect may increase somewhat as a result of the additional variation in weights arising from non-response adjustments. In computing the sampling fractions to be used to generate the required sample sizes, allowance needs to be made for non-coverage. With a 90 per cent coverage rate, sampling fractions need to be increased by 11 per cent.

E. Concluding remarks

75. An understanding of design effects and their components is valuable in developing sample designs for new surveys. For example:

- The magnitudes of the overall design effects for key survey estimates may be used in determining the required sample size. The sample size needed to give the specified level of precision for each key estimate may be computed for an unrestricted sample, and this sample size may then be multiplied by the estimate's design effect to give the required sample size for that estimate with the complex sample design. The final sample size may then be chosen by examining the required sample sizes for each of the estimates (perhaps, with the largest of these sample sizes being taken).

- When a disproportionate stratified sample design is to be used to provide domain estimates of required levels of precision, the resultant loss of precision for estimates for the total sample and for subgroups that cut across the domains can be assessed by computing the design effect due to variable weights. If the loss is found to be too great, then a change in the domain requirements that leads to less variable weights may be indicated.

- If the design effect from clustering is very large for some key survey estimates, then the possibility of increasing the number of sampled PSUs (a) with a smaller subsample size (b) should be considered.

76. While the formulas presented in this chapter are useful in sample design, they should not be applied uncritically. As noted in several places, the formulae are derived under a number of assumptions and simplifications. Users need to be sensitive to these features and to consider whether the formulae will provide reasonable approximations for their situation.

77. Estimating design effects from clustering requires estimates of ρ values for the key survey variables. These estimates are inevitably imperfect, but reasonable estimates may suffice. To err in the direction of the use of a value of ρ larger than predicted leads to the specification of a larger required sample size; hence, this is a conservative strategy.

78. Finally, it should be noted that the purpose of using these design effect models is to produce an efficient sample design. The failure of the models to hold exactly will result in some loss of efficiency. However, the use of inappropriate models to develop the sample design does not affect the validity of the survey estimates. With probability sampling, the survey estimates remain valid estimates of the population parameters.

References

Brick, J.M., and G. Kalton (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, vol. 5, pp. 215-238.

Bye, B., and S. Gallicchio (1989). A note on sampling variance estimates for Social Security program participants from the Survey of Income and Program Participation. *United States Social Security Bulletin*, vol. 51, no. 10, pp. 4-21.

Clark, R.G., and D.G. Steel (2002). The effect of using household as a sampling unit. *International Statistical Review*, vol. 70, pp. 289-314.

Cochran, W.G. (1977). *Sampling Techniques,* 3rd ed. New York: Wiley.

Gabler, S., S. Haeder and P. Lahiri (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology,* vol. 25, pp. 105-106.

Holt, D. H. (1980). Discussion of the paper by Verma, V., C. Scott and C. O'Muircheartaigh: sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society,* Series *A*, vol. 143, pp. 468-469.

Kalton, G. (1977). Practical methods for estimating survey sampling errors. *Bulletin of the International Statistical Institute*, vol. 47, No. 3, pp. 495-514.

_____ (1979). Ultimate cluster sampling. *Journal of the Royal Statistical Society,* Series *A*, vol. 142, pp. 210-222.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

_____ (1976). Optima and proxima in linear sample designs. *Journal of the Royal Statistical Society*, Series A, vol. 139, pp. 80-95.

_____ (1982). Design effect. In *Encyclopedia of Statistical Sciences*, vol. 2, S. Kotz and N.L. Johnson, eds., New York: Wiley, pp. 347-348.

_____ (1988). Multi-purpose sample designs. *Survey Methodology*, vol. 14, pp. 19-32.

_____ (1992). Weighting for unequal P_i. *Journal of Official Statistics*, vol. 8, pp. 183-200.

_____ (1995). Methods for design effects. *Journal of Official Statistics*, vol. 11, pp. 55-77.

_____, and M.R. Frankel (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, Series B, vol. 36, pp. 1-37.

_____, and others (1995). Design effects for correlated $(p_i - p_j)$. *Survey Methodology*, vol. 21, pp. 117-124.

_____, and others (1976). *Sampling Errors in Fertility Surveys*. World Fertility Survey Occasional Paper, No. 17. The Hague: International Statistical Institute.

Korn, E.L., and B.I. Graubard (1999). *Analysis of Health Surveys*. New York: Wiley.

Lehtonen, R., and E.J. Pahkinen (1994). *Practical Methods for Design and Analysis of Complex Surveys*, revised ed. Chichester, United Kingdom: Wiley.

Lepkowski, J.M., and J. Bowles (1996). Sampling error software for personal computers. *Survey Statistician*, vol. 35, pp. 10-17.

Rust, K.F. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, vol.1, pp. 381-397.

_____ , and J.N.K. Rao (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, vol. 5, pp. 283-310.

Skinner, C.J., D. Holt and T.M.F. Smith, eds. (1989). *Analysis of Complex Surveys*. Chichester, United Kingdom: Wiley.

Spencer, B.D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilities. *Survey Methodology*, vol. 26, pp. 137-138.

United Nations (1993). *National Household Survey Capability Programme: Sampling Errors in Household Surveys*. UNFPA/UN/INT-92-P80-15E. New York: United Nations Statistics Division. Publication prepared by Vijay Verma.

Verma, V., and T. Lê (1996). An analysis of sampling errors for the Demographic and Health Surveys. *International Statistical Review*, vol. 64, pp. 265-294.

Verma, V., C. Scott and C. O'Muircheartaigh (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society,* Series A, vol. 143, pp. 431-473.

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