Survey Methodology, June 1988
Vol. 14, No. 1, pp. 19–32
Statistics Canada

Multipurpose Sample Designs[1]

LESLIE KISH[2]

ABSTRACT

Most surveys have many purposes and a hierarchy of six levels is proposed here. Yet most theory and textbooks are based on unipurpose theory, in order to avoid the complexity and conflicts of multipurpose designs. Ten areas of conflict between purposes are shown, then problems and solutions are advanced for each. Compromises and joint solutions fortunately are feasible, because most optima are very flat; also because most "requirements" for precision are actually very flexible. To state and to face the many purposes are preferable to the common practice of hiding behind some artificially picked single purpose; and they have also become more feasible with modern computers.

KEY WORDS: Allocations to domains; Mean-Square-Errors; Multipurpose allocation; Multipurpose design; Optimal allocation; Periodic samples; Sample size.

1. INTRODUCTION

Most studies involve several purposes during the planning stages and then typically many more purposes emerge later during the analyses of data and more during their interpretation and utilization. However, the real multipurpose nature of most studies tend to remain hidden under the surface of oversimplified, univariate discussions of study designs. This seems most clearly evident for sample surveys, which I shall discuss here; but I believe that this discrepancy also holds for other statistical designs, such as experimental and evaluation studies.

In practice, surveys are usually multipurpose. Why then are multipurpose designs neglected in sampling theory? Because multipurpose theory would be too complex and difficult, and sampling theory is rather complex already; specific exceptions will be noted later. Even the descriptions we read of actual sample designs tend to follow and to borrow the prestige of univariate and unipurpose sampling theory, rather than to portray faithfully the many compromises of complex reality. Many common designs (especially equal probability of selection method) probably serve robustly a variety of purposes, *explicit* planning of multipurpose designs seems to be rare, though much needed, I propose.

There are several aspects to the *multipurpose nature* of survey samples, and these are displayed in a hierarchy of *six levels* in Section 2. Then *ten areas of conflict* between purposes are specified in Section 3. Sections 4 to 9 deal with specific areas of conflict, presenting approaches to and solutions for them. Some of these solutions are attributed to widely dispersed articles of survey sampling; but others are more novel, hence less fully developed, derived, and referenced.

In this overall review I aim first and foremost to serve practitioners with handy references on approaches, methods and procedures for multipurpose designs; to alert them both to the importance and to the feasibility of such designs. Second, I also wish to provide a framework for integrated, theoretical future work on the many problems and conflicts of multipurpose designs. Imperfections of my methods can serve as stimuli to others for better derivations for them, as well as for developing new methods.

---

[1] Keynote address at the International Symposium on Statistics, Taipei, Taiwan, August 1986; and also at a seminar of Statistics Canada, October 7, 1987.

[2] Leslie Kish, Institute for Social Research. The University of Michigan, Ann Arbor, MI 48104 USA.

2. A HIERARCHY FOR LEVELS OF PURPOSES

To begin with, we need some clarification of the meaning of "multipurpose", because too many concepts are confused under this term in our literature. To reduce the confusion, I classified a score of purposes into six levels in Table 1. Most of the time either multiple variables or multi-subject surveys (levels 3 or 4 in Table 1) are discussed and "multi-subject" (4) has sometimes been distinguished from multipurpose (3) for the same or closely related variables (Murthy 1967). Each of these six levels is shown in several specific manifestations, which can be usefully augmented and discussed in more detail elsewhere (e.g., United Nations 1980; Lahiri 1963).

Integrated survey operations on level 5 are related to, but should be distinguished from multi-subject surveys, because they refer to organizations and institutions that conduct many surveys in diverse fields over longer periods of time (United Nations 1980; Foreman 1983). An earlier name was "continuing survey operations", when it was recognized that most large-scale, wide-spread sample surveys were conducted by continuing survey organizations like the U.S. Census Bureau, Statistics Canada, or our Survey Research Center. Such continuity has large advantages in costs and quality, with restraining effects on sample designs (Kish 1965).

Master frames or master samples on level 6 refer to further extensions and specializations of multipurpose approaches. They may refer simply to using the same maps, or block listings, or area segments for several different surveys; or to the large-scale example of the "Master Sample of Agriculture" (King and Jessen 1945), where rural areas on the maps of all the counties of the USA were divided into segments of about four farms each; or to the firm that sells current listings of dwellings for most samples used in Western Germany. These very diverse examples have common bases in the savings from sharing the "startup" costs (of design, stratification, listing, etc.) for constructing sampling frames.

Diverse statistics based on single variable and diverse domains (levels 1 and 2) have been typically neglected in the literature of multipurpose sampling, although they are the most common, but they can have the most drastic effects and cause the most dramatic conflicts, as we shall see later. The effect of designs can be very different for statistics like medians and quantiles or regression coefficients than the effects for means and for aggregates (Kish 1961; Kish 1965; Kish and Frankel 1974). Furthermore, designing for period samples brings on new considerations (Section 8). But most dramatic effects can be seen simply for the means of small "subclasses" (e.g., as small as 0.10 or 0.01) of the entire sample, representing similar "domains" in the population (Section 5).

Each of the six levels of purposes presents different aspects for designs and each level can be fruitfully explored for more specific meanings and examples, some of which are listed in Table 1.

The difficulties of multipurpose designs, which have caused them to be neglected and avoided, are of several kinds. First, the different purposes must be formulated *explicitly in statistical terms*, so that these may serve in formulas for their comparisons and for formulated compromises; but obtaining a (complete) list of such explicit, formal terms may be the principal obstacle. Second, estimates of *variance and cost factors* are needed for each purpose. Third, for some methods values must be obtained for the assigned to the *"required" precisions* for all the purposes (Section 5). Fourth, the above values and estimates must be combined in a mathematical formulation in order to arrive at the *solution of a single "optimal" design* to be actually used. The computational tasks for such solutions have been eased by electronic computers, but the conceptual and theoretical tasks remain (Section 5).

The difficulties of these tasks help to explain why discussions of multipurpose designs have been largely neglected designs in textbooks. However, note later references and bibliography here and in Rodriquez-Vera (1982); also Cochran (1977), and Chatterjee (1967). Furthermore, also in descriptions of actual surveys, often a single statistic (e.g. the mean) of a single principal variable is presented as *the* only (principal) purpose for the study. In the framework of multipurpose design

Table 1

Hierarchy of Purposes

1. Diverse statistics from the same variables
   - Totals or means or medians and quantiles, distributions
   - Analytical statistics: regressions, categorical analysis
   - Time aspects: static, macro-change, micro-change, cumulative
2. Diverse populations and domains (subclasses)
   - Proper classes and crossclasses
   - Comparisons of subclasses
3. Multiple variables on the same subject
   - Alternative measures of one variable; e.g. of income, or unemployment
   - Diverse periods — per day, week, month, year
   - Several aspects of one subject: income, savings, wealth
4. Multisubject surveys
   - Several subjects on same schedule, interview, operation
   - Health surveys of many diseases
   - Market research for several clients, many goods
   - Agricultural surveys of many crops
   - "Omnibus" social surveys
5. Continuing, integrated survey operations
   - NSS in India, CPS in USA, NHSCP of UN
   - Separate surveys from one office and field staff
   - Common source of surveys
   - Diverse methods, costs, operations, allocations, respondents
6. Master frames
   - Several samples from one frame or set of listings
   - Separate institutions, organizations
   - Separate field staffs? Same PSU's?

design this is equivalent to assigning zero importance to all other purposes. The unreality of this pretense may be softened by assuming that other principal purposes would result in similar allocations; but this pretense should be buttressed with calculations of the four steps above.

3. AN OVERALL VIEW OF TEN AREAS OF CONFLICT

A brief overall view of ten areas of conflict, listed in Table 2, should be useful before we look at specific problems and possible solutions for each. The list will probably not prove exhaustive, and readers may well find other areas. Even more likely, they may find within these ten areas other problems and other solutions not explored here. It would be convenient if the ten areas of conflict should be linked rationally to the twenty purposes presented in six levels; we then could reduce this presentation to say, twenty purpose/conflict nodes or to ten level/conflict nodes. Unfortunately the areas of conflict denote a perpendicular dimension to the purpose and all (or most) of the 10×6 cells have meaningful contents.

Of this long list of ten areas of conflict fortunately not all need to be formulated for every actual sample design. I believe that possible conflicts about a) the sample sizes m_g and about b) the relation of biases to sampling errors should always be considered, at least informally, because they

are ubiquitous. Also c) allocation among domains and d) allocation among strata should receive at least a brief discussion, and often more. Computing sampling errors (j) should also be done on most surveys. However, in the common case of one-time surveys, conflicts i) about design over time need not be considered. On the other hand, in a continuing operation with a continuing sampling frame, the decisions about e), f), g), and h) (stratification, cluster sizes and measures) may have been made a long time ago for a fixed design. However, the cluster sizes (f) used in intermediate stages (blocks and segments) may be open to flexible operational changes.

It is also reassuring to know that compromises based on statistical methods can yield quite acceptable results, for several reasons (Sections 5-8). First, because moderate departures from optimal allocation result in only small or negligible increases of variance. Curves of efficiency tend to be flat within broad areas around the optimal points; thus great accuracy for separate designs, which would not be feasible, are not needed. Second, because wide departures from optimal allocations can, on the other hand, cause moderate to large increases in variances. Thus, ignoring important purposes can result in substantial losses of efficiency for them, and therefore those purposes should be included in compromise designs. Third, compromise designs, in accord with statistical methods, can reduce drastically the potentially large losses from allocations optimized for other purposes, and with only small increases over the separate optimal designs for each purpose (Section 5).

4. SAMPLE SIZES AND BIAS RATIOS (B/σ)

These two areas of conflict, a and b in Table 2, should perhaps be considered most important overall, because they can be most dramatic. We treat them together here only because they may be closely related through the effects of subclasses. Let us begin with the familiar (simple random sampling with replacement) sample size $m = S^2/V^2$ needed to yield a "required" precision $= V^2$ for a sample mean $\bar{y}$, with element variance $= S^2$. However, the S_g^2 depend greatly on the variables and on the domains, indexed jointly with g for the year $\bar{y}_g$; and the "required" V_g^2 may vary even more. We also include design effects D_g^2 that also vary, and thus $m_g = S_g^2 D_g^2 / V_g^2$ expresses the sample size needed for the mean of the variable g. For the mean $\bar{y}_g$ of a domain g, comprising only the proportion P_g in the population the *overall* sample size needed for the domain becomes $n_g = m_g/P_g$, and it is more practical to formulate the needed sampling fraction $f_g = n_g/N = S_g^2 D_g^2 / V_g^2 P_g N$. The factor (1-f) may be neglected or included in D_g^2. The P_g become small and critical if high precisions are "required" for small subclasses.

For comparisons of subclasses the variances increase even more: $m_g = (m_a^{-1} + m_b^{-1})^{-1} = n(P_a^{-1} + P_b^{-1})^{-1}$, with the P_a and P_b denoting proportions in the sample n (assuming $S_a^2 = S_b^2$). E.g., for the comparison of two subclass means of $0.01n$ and $0.10n$, we have the "effective size" $m_g = n(0.01^{-1} + 0.10^{-1})^{-1} = n/110$. For other statistics, such as medians and regression coefficients, formulating "required" sample sizes would become complex. It is more than we may discuss here, but some numbers may probably be specified.

Considerations for subclass statistics become greatly modified if, in addition to variances σ^2, we also include biases B^2 in the Root-Mean-Square-Error = RMSE = $\sqrt{(\sigma^2 + B^2)}$ for measures of accuracy. Figure 1 is meant to portray a common tendency in the accuracy of survey data, although great differences in the relations of biases to sampling errors are possible; reading the legend is urged here. It occurs commonly that potential biases B_1 are greater than the measurable sampling and variable errors σ_1, for the entire sample. However, on the horizontal axis the standard error σ_1 is shown to increase by a factor of about 3 for σ_2 of a subclass of about 1/10 of the total sample. For comparisons (differences) of two such subclasses σ_3 increases by about 1.4 more.

Table 2

Ten Areas of Conflicts (a-j)

a.(4)[1] Sizes m_g or rates f_g are needed for purposes g

$V_g^2 = S_g^2 D_g^2 / m_g$ and $m_g = S_g^2 D_g^2 / V_g^2$ or $f_g = S_g^2 D_g^2 / V_g^2 P_g N$

where m_g denote subclass sizes and $f_g = n_g/N = m_g/P_g N$ denote sampling rates

b.(4) Relation of biases to sampling errors in RMSE $= \sqrt{(\sigma^2 + B^2)}$
- The bias ratio B/σ decreases as σ increases for subclasses
- For comparisons B/σ tends to be small as B decreases, σ increases

c.(5) Allocation of the m_g among domains

$m_t = \Sigma_g m_g$

d.(6) Allocation of m_{gh} among strata h

$m_g = \Sigma_h m_{gh}$

e.(6) Choice of variables for stratification
Multivariate stratification

f.(7) Optimal cluster sizes

$D_g^2 = [1 + \rho_g(\bar{b}_g - 1)]\bar{b}_g = P_g n_t / a$ for crossclasses

g.(7) Measures for cluster sizes

h.(7) Retaining sampling units (PSU's) for changed subjects, measures and strata and for diverse subjects.

i.(8) Design over time
How much overlap? Panels? Change versus cumulation.

j.(9) Computing and presenting sampling errors.

[1] The numbers (4) to (9) refer to sections with treatments.

However, the hypotenuses denoting the RMSE are shown to increase much less. In $RMSE_1$ the bias B_1 is shown to dominate, and this may happen for some variables in large total samples. However, the subclass $RMSE_2$, because the bias was kept constant at $B_2 = B_1$, increased only moderately and is dominated by σ_2. This is even more true for $RMSE_3$, where the σ_3 has increased, but the biases — assumed to have the same sign, because that is a common tendency — decrease B_3 in the difference of means.

Examples of these phenomena abound everywhere and for all purposes are listed in Table 1. We choose the best known, critical statistics of unemployment, where admitted measurement biases may completely swamp the low values (e.g., 0.1 percent) of measurable fluctuations. However, for small subclasses (e.g. Black teenage boys) the sampling errors for small sample bases overtake the biases. For periodic comparisons the sampling variations become even more critical.

These relations among biases and variable errors assumed here are not logically necessary, but empirical and common. Neglect of these simple relations leads to a great deal of confusion concerning the need for sample surveys of adequate precision, i.e. with small sampling errors, σ. I propose Figure 1 as practical answers to some common questions, such as: Why do we spend

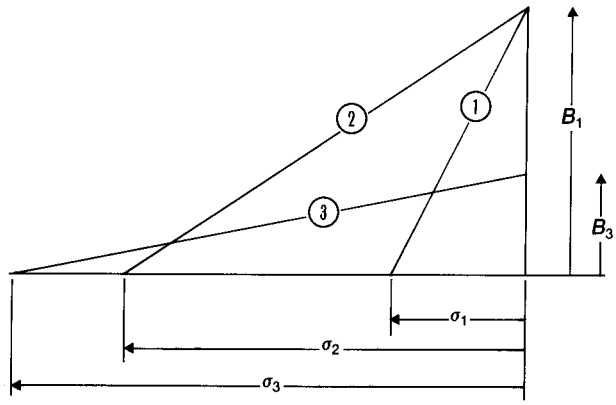

Figure 1. Variable errors (σ) and biases (B) in root mean square errors (RMSE)

The bases represent sampling errors and other variable errors (σ). For example σ_1 may be the $ste(\bar{y}_t)$ for the mean $\bar{y}$ of the entire sample and σ_2 may be a larger $ste(\bar{y}_c)$ for a subclass mean, and σ_3 may be the $ste(\bar{y}_c - \bar{y}_b)$ for the difference between two subclass means.

The heights represent biases (B) and the hypotenuse denotes the RMSE $= \sqrt{(\sigma^2 + B^2)}$. (1) For the entire sample the bias B_1 may be large compared with the variable error σ_1, thus taking larger samples would not decrease the $RMSE_1$ by much. (2) However with the same bias B_1, but with a smaller sample in the subclass, the ratio changes and the σ_2 dominates the $RMSE_2$; and this is not much larger than for (1) despite a much smaller sample. (3) Furthermore, for the difference of means, the net bias B_3 may be much smaller; so that even with a larger σ_3, the $RMSE_3$ for the difference is but little greater than $RMSE_2$. This drastic change in the bias ratio B/σ tends to appear not only for differences between subclasses within the same sample, but also for differences between repeat surveys.

money for large samples and on rigorous sampling methods in the face of large measurement biases? Why bother computing sampling errors when response biases dominate the total error? The implicit answers come from the domination of sampling errors in the subclasses, and even more in their comparisons. Let us make these implicit answers more explicit in future sample designs.

5. ALLOCATION AMONG DOMAINS

This most important and frequent area of conflict has several aspects. First, consider the allocation of total sample size (or effort or cost) among the domains that constitute a partition of the total population. A common example is allocation among the several (5, 10, 20 or 50) provinces or regions or states of a country; those domains typically have very unequal populations N_d, with ranges of 1 to 100 perhaps in relative sizes, though they may cover roughly equal surface areas. Often the question takes this form: Should the sample sizes n_d be roughly equal; or should the n_d be proportional to the N_d, with constant sampling rates $f_d = f$? Equal n_d tends to yield roughly equal errors, $ste(\bar{y}_d)$ for the means. On the other hand, constant $f_d = f$ tends to yield the lowest $ste(\bar{y}_w)$ for the overall mean $\bar{y}_w = \Sigma W_d \bar{y}_d$, because it yields lower errors for the larger domains. This error may be lower than "required" for $\bar{y}_w$, especially in view of potential biases (Figure 1), and may not justify large total sample sizes and costs. This is the contention of proponents of equal sizes n_d for provinces. However, increased sampling errors for $\bar{y}_w$ are also suffered by most other subclasses, especially "crossclasses" like age, sex, socioeconomic classes, etc. whose sizes tend to proportionality to the total. Those are common disadvantages of the highly unequal $f_d = n_d/N_d$ for provinces that result from the equal n_d values.

For example, in the Current Population Surveys of the USA, larger f_d are assigned to the smaller states. The resulting weighting increases the variances (for a fixed total cost) of the overall means and also of "crossclasses", such as young men and women, and especially of Black teenage

boys and girls (with critically high unemployment rates). Similar conflicts between national and provincial needs occur in all countries, because provinces have widely different populations. The need for better provincial data, for fixed total cost, conflicts with greater precision for national and for "crossclass" statistics.

To reduce the usual confusion, I distinguish "domains" to denote partitions of the population, from "subclasses," the corresponding partitions of the sample. Then I distinguish "design domains" (and subclasses) to refer to partitions (like provinces and regions) that are contained in strata defined by the sample design, from "crossclasses" (like age, sex, occupation, income, etc.) that cut across the sample design, both clusters and strata, often almost randomly. The design effects differ for these two types of subclasses (Kish 1961, 1980, 1987).

In addition, other sources of conflict may arise from *domain* differences other than their sizes: in the distribution of variables, also in the variances $D_d^2 S_d^2$ precisions; but we need not enter into those complexities here. Beyond calling attention to the problems, we refer to two distinct technical methods for the joint solution of the conflicts in allocation, (the fourth step noted at the end of Section 2). One approach uses iterative nonlinear programming in order to satisfy for *minimal cost* the "required" precisions jointly for all stated purposes. These elegant solutions to diverse problems exploit modern computers and have been published in many articles since 1963 (see reviews and references in Bean and Burmeister 1978, Rodriquez-Vera 1982, Cochran 1977). The "required minimal" cost often turns out much too high, because the "required" precisions were unrealistic. Then the solutions are drastically rescaled downwards. But such rescaling exposes the false pretensions (in my view) of this elegant approach that depends on unrealistic "required" precisions. Principally, I question the reality of "step functions" for "required" precisions that assign a constant value to any variance below the required V^2 and zero value to variances above it.

A very different approach calls for some form of *averaging* between all the "optimal" (preferred) allocations for various purposes, by *minimizing the combined* (weighted) *variance* either for fixed cost or fixed sample size. Of course, if the resulting combined variances turns out to be too high (or low), the solutions can be scaled up (or down) in total fixed cost or sample size. I prefer this solution, which compromises between different allocations, each of which would optimize for only one purpose (Yates 1981; Dalenius 1957). It involves assigning relative values of importance I_g to all the list statistics and this may seem difficult (but an "ignorant" decision-maker can assign equal I_g to all of them). But the other two alternatives are more extreme and they are bound to prove even more difficult: either to specify the "required" precisions of all statistics for the first approach, which then assigns arbitrarily equal weights of importance to all of them; or to specify one statistic for the total weight of one, and thus zero weights for all other statistics.

Furthermore, compromises for the average can be shown to be generally feasible and worthwhile, because the allocations are insensitive to moderate changes of weights of important (as is often true in statistics). After all, changing the relative importance by ratios of e.g., 2 or 5 should be less drastic than assigning the total weight 1 to one variable and 0 to all others, a process that implies infinite ratios of importance.

First, denote with $\Sigma_i V_{gi}^2/n_i$ the variance attainable for a statistic g with the allocations of sample sizes n_i for the ith component of variation. Then let $1 + L_g(n) = (\Sigma_i V_{gi}^2/n_i)/V_g^2(\min) = \Sigma_i C_{gi}^2/n_i$ denote the ratio of increase (with the allocation n_i) in the variance of the gth statistic over its own minimal variance, both for the same fixed Σn_i. Thus $L_g(n)$ is the *relative* loss over the minimal value of 1, and accepting the relative variances C_{gi}^2/n_i as the functions to be minimized is a critical decision; those functions seem to me more reasonable than any others that I can imagine for the functions to be combined in (1) below. For example, I prefer them to the V_{gi}^2 which depend on arbitrary units of measurement, which are removed by the $V_{gi}^2(\min)$. But in rare cases we may be faced with $V_g^2(\min) = 0$ or very small and this may make C_{gi}^2 widly large and unstable; in these

Table 3

Loss functions (1 + L) for two populations (Kish 1976)

Allocations m_i	(A) (1 + L) for $W_2/W_2 = 4$			(B) (1 + L) for 133 countries: 0.2 to 100 mm			
						Joint with weights	
	$\Sigma W_i \bar{y}_i$	$\Sigma \bar{y}_i/2$	Joint	$\Sigma W_i \bar{y}_i$	$\Sigma \bar{y}_i/133$	1:1	I_c/I_d:1
MW_i	1	1.56	1.28	1	6.86	3.93	
M/H	1.36	1	1.18	3.34	1	2.17	
$\propto \sqrt{(W_i}$	1.08	1.125	1.102	1.35	1.54	1.44	
$\propto \sqrt{(W_i^2 + H^{-2})}$	1.116	1.080	1.098	1.31	1.28	1.295	
$\propto \sqrt{(0.5 W_i^2 + H^{-2}}$				1.47	1.17	(1.32)	1.27
$\propto \sqrt{(2 W_i^2 + H^{-2})}$				1.20	1.44	(1.32)	1.28
$\propto \sqrt{(4 W_i^2 + H^{-2})}$				1.12	1.66	(1.39)	1.23

In (A) there are two strata and domains ($W_1 = 0.8$ and $W_2 = 0.2$); note that the allocation $m_i = \sqrt{W_i}$ does almost as well for the joint loss as the optimal.

In (B) we have the populations of 133 countries, ranging in size from 0.2 to over 100 millions, a range of 500 in relative sizes. From this problem of allocation (for the World Fertility Survey) we omitted, for practical reasons, the four largest countries and a few under 0.2 millions. Their inclusion would raise the variance of relative sizes, W_i, from 2.5 to 12, and would make the results more dramatic. Note that the $\sqrt{W_i}$ allocation reduces losses quite well. Some compromise is better than none. But the optimal allocation, $\sqrt{(W_i^2 + H^{-2})}$, is considerably better. Different values of I_c/I_d(= 1/2, 2/1 and 4/1) increase slightly the variance of the joint loss function with (1:1) weights; but they remain steady for joint loss functions with their own weights I_c/I_d:1.

Two examples in Table 3 illustrate the surprisingly good compromises between conflicting allocations yielded by the method of weighted averaging: its results on the fourth row of Table 3 compare very favorably with the others. The reasons for the excellent results come from the very broad flat surfaces for the optimal allocations, as discussed in Section 2 and shown elsewhere (Kish 1976; Kish 1987). For example, in Canada the 10 provinces vary seventy-fold from smallest to largest population sizes, and thus resemble B in Table 3; they serve as a graphical illustration in Figure 2. (See also Fellegi and Sunter 1974.)

cases assign arbitrary values to the C_{gi}^2 or to the I_g below. These and the following including Table 3 are developed and discussed by Kish (1976).

Then with the weights I_g assigned for relative importance of the gth statistic for any set of allocations n_i of the sample sizes,

$$1 + L(n) = \Sigma_g I_g (1 + L_g(n)) = \Sigma_g I_g \Sigma_i C_{gi}^2 / n_i$$

$$= \Sigma_i \Sigma_g I_g C_{gi}^2 / n_i = \Sigma_i Z_i^2 / n_i. \qquad (1)$$

After changing the order of summation, we created the new variables $Z^2{}_i = \Sigma_g I_g C_{gi}^2$. This function may be minimized to give compromise solutions for fixed total cost $\Sigma c_i n_i$. For the conflict between $n_d = n/H$ of equal sample sizes for domains versus $n_d = nW_d$ proportional to domain sizes W_d, the optimal compromise allocations are found to be proportional to $\sqrt{(W_d^2 + H^{-2})}$, with equal values for I_g.

An important example was provided by the (otherwise excellent) World Fertility Surveys, which used roughly equal sample sizes for small and large countries: actual sample sizes varied only within the range of 3 to 10 thousand and with no discernible correlation with population size. Consequently, there were two- or three-fold increases of variances in the continental averages of national surveys, their "main contributions to knowledge":

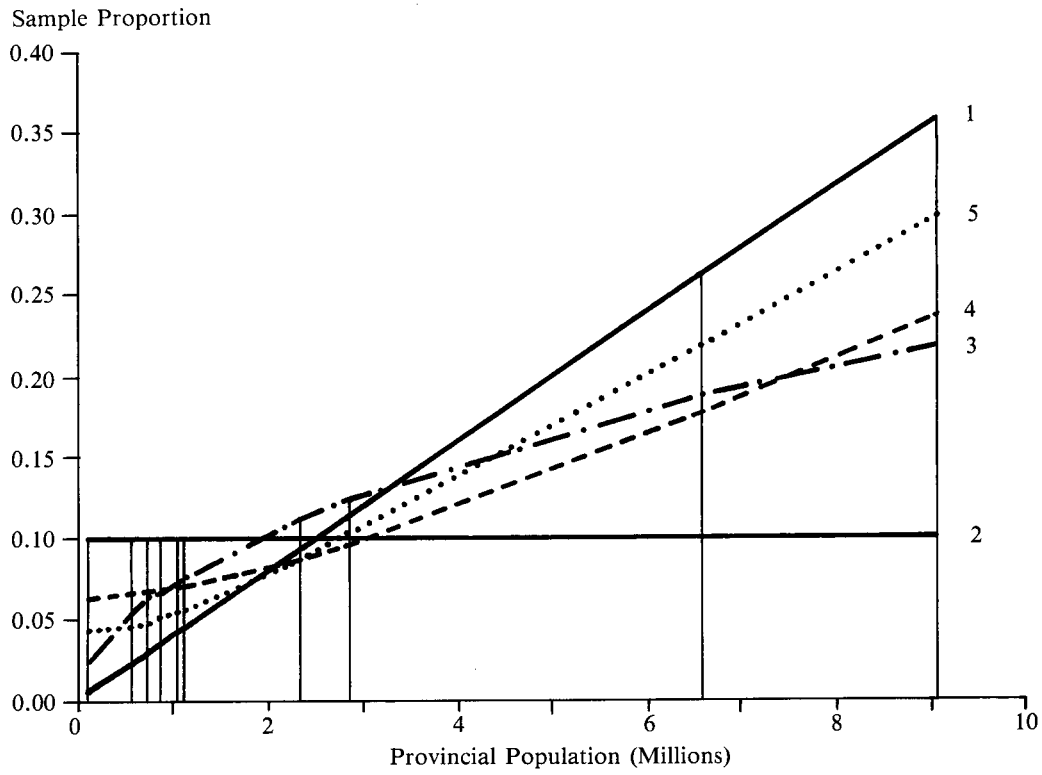

Figure 2. Five Alternative Allocations of Sample Sizes n_h of Fixed Total Σn_h

The ten provinces of Canada illustrate graphically the usual conflicts from major domains with unequal sizes, also the feasible successful compromises.

1 Allocation proportional to domain sizes $n_h \propto W_h$ is diagonal.
2 Equal allocation $n_h \alpha 1/H$ is a horizontal.
Divergences of the two allocations are large near the ends.
3 The square-root allocation, $n_h \propto \sqrt{W_h}$ yields compromises at both ends.
4 The "optimal" allocation $n_h \propto \sqrt{(W^2 + 1/H^2)}$ improves both ends, and especially with an appealing "floor" near the lower end.
5 A "weighted" optimal $n_h \propto \sqrt{(.8W^2 + .2/H^2)}$ improves the upper end considerably.

"So far, the main contribution to knowledge has been to confirm the downward trend in fertility that characterized much of Asia and Latin America in the 1970's and to highlight the contrast with Africa where both fertility and the desire for large numbers of children remain high" (Macura and Cleland 1985).

6. ALLOCATIONS TO STRATA AND CHOICE OF STRATIFIERS

Domains and strata often get confused in discussions, but the two aspects should be kept distinct in practical work on designs. Domains refer to subpopulations for which separate estimates are sought, whereas strata are usually smaller partitions created for decreasing variances. For example, within provinces as domains more strata may be created to reduce province variances; but cross-domains like age, sex and economic status tend to straddle across the strata. Allocations of sample sizes to strata, though often not as crucial as allocations to domains, may be important in case

of efficient disproportionate optimal allocations. The two methods of Section 5 for allocating sample sizes to domains can also be applied to allocations to strata, although the aims differ. Some of the references on nonlinear programming refer to domains and others to strata, and some confuse the two.

The presence of several survey variables and statistics among the purposes have clear implications for using more stratifying variables. Different survey variables will tend to have diverse optimal relations with the stratifiers; then it is best to use many stratifiers, even if each stratifier is used with only few stratum divisions (categories). Multipurpose design is the best reason for multivariate stratifiction (Kish and Anderson 1978). It may also best justify the need for "controlled selection" methods. The choice of stratum boundaries, called "optimal stratification", is a related topic, but of less importance in this condensed presentation.

7. CLUSTER SIZES; MEASURES OF SIZE; RETAINING UNITS

In descriptions of sample designs we find sometimes that the design effect has been approximated with $D_g^2 = [1 + \rho_g(\bar{b}_t - 1)]$, where ρ stands for a synthetic intraclass correlation of the "most important" variable g and $\bar{b}_t = n/a$, the average cluster size. This would yield the effective element variance $S_g^2 D_g^2$ and the variance $S_g^2 D_g^2/n$ for the mean of the variable g. However, we must question the contents of n and of $\bar{b}_t$. If our population consists of married women of childbearing age, they may be only 10 percent of total persons and found in only 30 percent of dwellings; and much fewer than that for some rare populations. This situation has been treated in sampling for rare traits (Kish 1965). "Ordinarily we avoid large clusters, because of their adverse effects on the variance. But even large clusters of the entire population will yield only small clusters of a rare trait, if this is widely spread. For example, entire blocks may be sampled for persons over 65 years of age; entire villages may be searched for persons with an identifiable disease. If, on the contrary, the trait is concentrated in small areas, those areas often can be recognized and stratified accordingly."

In multipurpose designs, the crossclasses of the sample will be of variable sizes that are portions of the total sample size n_t, with $\bar{M}_g$ as their different proportions in the populations. Thus we want to estimate in the design not only $[1 + \rho_g(\bar{b}_t - 1)]$ for diverse variables g for the total sample n_t, but also $[1 + \rho_g(\bar{b}_t - 1)]$ for many crossclasses. Here, as in Section 6, the index g is made to serve both variables and subclasses, in order to simplify notation. Then we make use of some conjectures that have been shown to be good approximations in thousands of empirical computations for scores of samples:

$$[1 + \rho_g(\bar{b}_g - 1)] \simeq [1 + \rho_g(\bar{M}_g \bar{b}_t - 1)] \simeq [1 + \rho_t(\bar{M}_g \bar{b}_t - 1)] \qquad (2)$$

That is, we use $\bar{b}_g = \bar{M}_g \bar{b}_t$ and $\rho_g \simeq \rho_t$ as rough approximations. True that this somewhat underestimates the average values of D_g^2 for crossclasses, because of variations in cluster sizes of crossclasses. But that is a small factor compared to the large variations of ρ_g between variables (Kish 1987; Verma *et al.* 1980; Kish *et al.* 1976), and that underestimate has small effects on the efficiency of designs. It is important to consider efficiencies of estimates for subclasses as well as for the entire sample; these considerations point to considerably higher efficiencies for larger clusters than would be shown for $\bar{b}_t$ and n_t for the total sample only.

Measures of size are related to cluster sizes, but differ because of errors in the available measures, due especially to different population contents and to obsolescence. We must also note problems concerning measures of size for multisubject surveys and for "integrated survey operations" for

different populations, which may especially need drastic compromises. Those two levels of purposes (Table 1) should be distinguished because multisubject surveys use single samples in one operation; but integrated survey operations may use different sizes of sampling units for different surveys (United Nations 1980). For example, consider integrated designs for total populations and for agriculture; also perhaps for ethnic subpopulations; also perhaps for industrial or business activities: the measures of size for each of these may differ greatly. Yet some compromise solution may be found to yield reasonable efficiencies for each.

Measures of size are also closely related to problems for "Retaining units after changing strata and probabilities" (Kish and Scott 1971). Those methods were designed to deal with changes over time of sampling units, both in measures of size and in stratifying variables; but the methods are also relevant for differences between survey variables:

"Unequal selection probabilities are often assigned to sampling units. Our methods, though more generally applicable, are especially needed for the selection of primary sampling units for surveys. Often these are selected separately from many strata, with one selection from each stratum.

"After the initial selection the units may be used for many surveys over several years. But as time passes, the needs of new surveys may be better served by new strata and new selection probabilities, based on new data, than by those used for the initial selection. The difference between initial and new data may be due to differential changes among the sampling units as revealed by the latest Census. Or the differences may be due to changes in survey objectives and populations; for example, a sample initially designed for households and persons may later be required to serve a survey of farmers, or college students. *Obviously our methods are also applicable to designing simultaneously a related group of samples with differing objectives.*"

This method allows for using the best measures (for size and for strata) separately for each sample purpose, but maximizing the retention of the overlap of sampling units between the samples for separate purposes (especially PSU's). However, it would be possible to design a compromise that would average the measures in order to achieve a complete overlap of units, but sacrificing some efficiency for each of the purposes. A compromise between the two techniques may be even better than either: increase the overlap with small sacrifices of separate efficiencies by recognizing only differences of measures that surpass some arbitrary minimal criteria (Kish and Scott 1971).

8. PURPOSES AND DESIGNS FOR PERIODIC STUDIES

Periodic studies provide areas of conflict with great and growing importance as their numbers and sizes increase. It is wrong to assume that those expensive and influential surveys have only one of the five purposes listed in Table 4, because usually they are needed for several or all, if the design permits their use.

In Table 4 we note five purposes and six designs. The first four are paired with similar letters on the same four lines. These pairings call attention to designs that best serve, with reduced variances, each of the four purposes. Most periodic studies have several purposes and thus we should face, and perhaps solve, the difficult problems of multipurpose designs. Actually current levels (A) and net changes (C) can be served with any of the six listed designs, but with some increase in variances or in costs. However, individual (gross, micro) changes (D) need panels; and cumulations (B) need some changes of samples, and are fastest without any overlaps. For current levels (A) variances can be somewhat reduced with estimators using correlations from partial overlaps. Net changes (C) benefit from correlations from any overlap, and most from complete overlaps (Cochran 1977; Kish 1987; Kish 1965). Reasonable compromises often become possible, when purposes can be defined. However, extraneous considerations may rule out some designs (e.g., overlaps may be either prohibited or enforced) and thus force the use of less efficient — but still valid — designs.

Table 4

Purposes and Designs for Periodic Samples

Purposes	Designs	Rotation Scheme
A. Current levels	A. Partial overlaps $0 < P < 1$	abc-cde-efg
B. Cumulations	B. Nonoverlaps P = 0	aaa-bbb-ccc
C. Net changes (means)	C. Complete overlaps P = 1	aaa-aaa-aaa
D. Gross changes (individual)	D. Panels	same elements
E. Multipurpose time series	E. Combinations, SPD	
	F. Master Frames	

The chief variation in these six designs concerns the amount (and kind) of overlaps between periods. The rotation scheme of complete overlaps shows, with aaa-aaa, that the periods have all common parts; the nonoverlap with aaa-bbb shows none; and the partial overlap abc-cde-efg shows c and e as 1/3 overlaps between succeeding periods only. This section concentrates on the effects of varying proportions of overlaps P in diverse designs on different purposes; in complete overlaps $P = 1$, in nonoverlaps $P = 0$, and in partial overlaps $0 < P < 1$. The purposes are discussed in terms of variances for estimated means, because means (and percentages, rates, proportions) are both the most used and the simplest estimates. Effects on other estimates will not be entirely different but they are too many, diverse, and difficult to be explored here.

More discussions of panels is also available elsewhere, with its advantages, disadvantages, problems and solutions (Duncan and Kalton 1986; Kish 1987). I call attention to SPD, or Split Panel Designs, that I am trying to promote for multipurpose designs. These would combine a panel sample P with new rotating or "rolling" samples, so that *Pa-Pb-Pc-Pd* would symbolize the periodic samples. The rolling samples a,b,c,d etc., could be cumulated into larger samples. The panel P serves primarily to provide micro (individual gross changes). But it also serves as the partial overlap for better estimates of both current levels and macro (mean, net) changes *for any pair of periods.*

9. COMPUTING AND PRESENTING SAMPLING ERRORS

It seems questionable to include this topic under design, but I have no doubt that this is a multipurpose problem. The strategies for computing and presenting sampling errors deserve separate listing as an area of conflict among the many statistics given generally for the results of surveys. It is not enough to present standard errors for only one or a few of the most important statistics: they are too many and too diverse. Because of that diversity, the practice has grown up to compute from the variances other expressions of sampling variability, especially estimates of the "design effects" d_g^2; also sometimes from the $d_g^2 = 1 + \rho_g(\bar{b}_g - 1)$, estimates of the synthetic intraclass correlation ρ_g.

Briefly, I advise: a) Compute sampling errors for many variables, because the variances, the design effects (d_g^2), and the intraclass coefficients (ρ_g) can and do differ greatly between variables. b) You may have to do some averaging of sampling errors, because it may be inconvenient or confusing to present them all. c) It may be neither feasible nor necessary to compute sampling errors for all subclasses, because they can often be approximated with reasonable models. d) It is necessary to present sampling errors for subclasses and for other statistics to guide the readers of the reports (Kish 1965; Kish 1987; Verma *et al* 1980). I hope that this topic will receive in the future from theorists and methodologists some of the attention it needs.

10. CONCLUSIONS

For the ten areas of conflict of Section 3 approaches and solutions are proposed in Sections 4 to 9 that are very diverse. Averaging allocations among domains in Section 5 seems to give surprisingly good compromise solutions. The advice in Section 6 to use more stratifiers can also yield worthwhile gains. In Sections 4 and 7 considerations for subclass estimates lead to drastically different decisions for sample designs. In Section 8 we note how periodic designs can be best suited to purposes, and best compromise for multipurpose aims. We looked at the different levels of purposes and at the various areas of conflicts jointly. Asking the right question is the core of most problems. I propose multipurpose design as a new paradigm, to replace "optimal" solutions to artificially partial questions such as: What is the optimal allocation for the mean $\bar{y}$ or the total $\hat{Y}$ of "the most important" variable?

REFERENCES

BEAN, J.A., and BURMEISTER, L.F. (1978). A review of optimal sample allocation for multipurpose surveys, *Biometrika*, 20, 3-14.

CHATTERJEE, S. (1967). A note on optimum stratification, *Skandinavisk Actuarietidskrift*, 50, 40-44.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed). New York: John Wiley and Sons.

DALENIUS, T. (1957). *Sampling in Sweden*, Stockholm: Almquist and Wicksell.

DUNCAN, G.J., and KALTON, G. (1986). Issues of design and analysis of surveys across time. *International Statistical Review*, 54.

FELLEGI, I.P., and SUNTER, A.B. (1974). Balance between different sources of survey errors. *Sankhyā*, 36, 119-142.

FOREMAN, E.K. (1983). Integrated programmes of household surveys: design aspects. *Bulletin of the International Statistical Institute.*

KING, A.J., and JESSEN, R.J. (1945). The master sample of agriculture, *Journal of the American Statistical Association*, 38-56.

KISH, L. (1987). *Statistical Design for Research*: New York: Wiley-Interscience.

KISH, L. (1986). Timing of surveys for public policy. *Australian Journal of Statistics*, 28, 1-12.

KISH, L. (1980). Design and estimation for domains. *The Statistician*, London, 29, 209-222.

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

KISH, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.

KISH, L. (1961). Efficient allocation for multipurpose samples. *Econometrica*, 29, 363-385.

KISH, L., and ANDERSON, D.W. (1978). Multivariate and multipurpose stratification, *Journal of the American Statistical Association*, 73, 24-34.

KISH, L., and FRANKEL, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*, B. 36, 1-37.

KISH, L., and SCOTT, A.J. (1971). Retaining units after changing strata and probabilities, *Journal of the Royal Statistical Society*, 66, 461-470.

KIREGYERA, B., and GACHUKI, P. (1985). Experiences in panel surveys: examples from an integrated sample survey programme in Kenya. *Bulletin of the International Statistical Institute.*

MACURA, M., and CLELAND, J. (1985). *A Celebration of Statistics: the ISI Centenary Volume*, (Eds. A.C. Atkinson and S.E. Fienberg), New York: Springer Verlag.

MURTHY, M.N. (1974). Evaluation of multi-subject sample survey systems. *International Statistical Review*, 42.

MURTHY, M.N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

UNITED NATIONS (1980). *National Household Survey Capability Programme* (NHSCP). New York: United Nations.

RODRIQUEZ-VERA, A. (1982). *Multipurpose Optimal Sample Allocation Using Mathematical Programming*. Ph.D. dissertation, The University of Michigan, Ann Arbor.

VERMA, V., SCOTT, C., and O'MUIRCHEARTAIGH, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society*, A, 143, 431-473.

YATES, F. (1981). *Sampling Methods for Censuses and Surveys*, (4th ed.). London: Griffin and Co.