

Article

Methods for oversampling rare subpopulations in social surveys

by Graham Kalton

December 2009



Methods for oversampling rare subpopulations in social surveys

Graham Kalton¹

Abstract

Surveys are frequently required to produce estimates for subpopulations, sometimes for a single subpopulation and sometimes for several subpopulations in addition to the total population. When membership of a rare subpopulation (or domain) can be determined from the sampling frame, selecting the required domain sample size is relatively straightforward. In this case the main issue is the extent of oversampling to employ when survey estimates are required for several domains and for the total population. Sampling and oversampling rare domains whose members cannot be identified in advance present a major challenge. A variety of methods has been used in this situation. In addition to large-scale screening, these methods include disproportionate stratified sampling, two-phase sampling, the use of multiple frames, multiplicity sampling, location sampling, panel surveys, and the use of multi-purpose surveys. This paper illustrates the application of these methods in a range of social surveys.

Key Words: Sample allocation; Screening; Disproportionate stratified sampling; Two-phase sampling; Multiple frames; Location sampling; Panel surveys; Multi-purpose surveys.

1. Introduction

I feel very privileged to have been invited to present this year's paper in the Waksberg Invited Paper Series, a series that honors Joe Waksberg for his numerous contributions to survey methodology. I was extremely fortunate to have had the opportunity to work with Joe at Westat for many years and, as did many others, I benefited greatly from that experience. When faced with an intractable sampling problem, Joe had a flair for turning the problem on its end and producing a workable solution. Since the problem often concerned the sampling of rare populations, I have chosen to review methods for sampling rare populations for this paper.

One of the major developments in survey research over the past several decades has been the continuously escalating demand for estimates for smaller and smaller subclasses (subpopulations) of the general population. This paper focuses on those subclasses – termed *domains* – that are planned for separate analysis at the sample design stage. Some examples of domains that have been taken into account in the sample designs of various surveys include a country's states or provinces, counties or districts; racial/ethnic minorities; households living in poverty; recent births; persons over 80 years of age; recent immigrants; gay men; drug users; and disabled persons. When the domains are small (also known as *rare populations*), the need to provide adequate sample sizes for domain analysis can create major challenges in sample design. This paper reviews the different probability sampling methods that are used to generate samples for estimating the characteristics of rare populations with required levels of precision. Sampling methods for estimating the size of a rare

population are not explicitly addressed, although similar methods are often applicable. However, capture-recapture and related methods are not addressed in this paper.

An important issue for sample design is whether the aim of a survey is to produce estimates for a single domain or many domains. Although much of the literature on the sampling of rare populations discusses sample designs for a single rare domain (*e.g.*, drug users), in practice surveys are often designed to produce estimates for many domains (*e.g.*, each of the provinces in a country or several racial/ethnic groups). The U.S. National Health and Nutrition Examination Survey (NHANES) is an example of a survey designed to produce estimates for many domains, in this case defined by age, sex, race/ethnicity and low-income status (Mohadjer and Curtin 2008). In sample designs that include many domains, the domains may be mutually exclusive (*e.g.*, provinces or the cells of the cross-classification of age group and race/ethnicity) or they may be intersecting (*e.g.*, domains defined separately by age group and by race/ethnicity).

The size of a domain is a key consideration. Kish (1987) proposed a classification of *major domains* of perhaps 10 percent or more of the total population, for which a general sample will usually produce reliable estimates; *minor domains* of 1 to 10 percent, for which the sampling methods in this paper are needed; *mini-domains* of 0.1 to 1 percent, estimates for which mostly require the use of statistical models; and *rare types* comprising less than 0.01 percent of the population, which generally cannot be handled by survey sampling methods. Many surveys aim to produce estimates for some major domains, some minor domains and occasionally even some mini-domains.

1. Graham Kalton, Westat, 1600 Research Blvd., Rockville, MD 20850, U.S.A. E-mail: grahamkalton@westat.com.

Since the sample sizes for most surveys are sufficient to produce estimates of reasonable precision for major domains, there is generally no need to adopt the kinds of oversampling procedures reviewed in this paper. However, there are some important design features that should be considered. It is, for example, valuable to take major domains into account in creating the strata for the survey. This consideration is of particular importance with geographically defined domains and multistage sampling. If a geographic domain is not made into a design stratum, the number of primary sampling units (PSUs) selected in that domain is a random variable; the sampled PSUs in strata that cut across the domain boundaries may or may not be in the domain, creating problems for domain estimation. It is also valuable to have a sizable number of sampled PSUs in each geographical domain in order to be able to compute direct variance estimates of reasonable precision, implying the need to spread the sample across a large number of PSUs. At the estimation stage, it is preferable, where possible, to apply nonresponse and noncoverage post-stratification-type adjustments at the domain rather than the national level. Singh, Gambino and Mantel (1994) and Marker (2001) discuss design issues and Rao (2003, pages 9-25) discusses estimation issues for major domains. Major domains will receive little attention in this paper.

At the other end of the size continuum, even with the use of special probability sampling methods, the sample sizes possible for most surveys are not large enough to produce standard design-based, or direct, estimates of characteristics for multiple domains when many of the domains are mini-domains or rare types. An obvious exception is a national population census, but censuses too have their limitations. Since they are conducted infrequently (in many countries only once a decade), their estimates are dated – a particular concern for mini-domains, which can experience rapid changes. Also, the content of a census must be severely limited in terms of the range of topics and depth of detail. Very large continuous surveys such as the American Community Survey (U.S. Census Bureau 2009a; Citro and Kalton 2007), the French rolling census (Durr 2005) and the German Microcensus (German Federal Statistical Office 2009) have been developed to address the need for more up-to-date data for small domains, but a restriction on content remains (although the content of the German Microcensus does vary over time). Other exceptions occur at the border between mini-domains and minor domains. For example, since 2007 the Canadian Community Health Survey has provided estimates on the health status of the populations of each of Canada's 121 health regions based on an annual survey of around 65,000 persons aged 12 and over, with the production of annual and biennial data files (Statistics Canada 2008). By combining the samples across multiple

years, researchers are able to produce estimates for rare populations of various types.

In general, however, the maximum sample size possible for a survey on a specific topic is not adequate to yield a large set of mini-domain estimates of acceptable precision. Yet policy makers are making increasing demands for local area data at the mini-domain level. This demand for estimates for mini-domains, mainly domains defined at least in part by geographical administrative units, is being addressed by the use of statistical modeling techniques, leading to model-dependent, indirect, small area estimates. Thus, for example, the U.S. Census Bureau's Small Area Income and Poverty Estimates program produces indirect estimates of income and poverty statistics for 3,141 counties and estimates of poor school-age children for around 15,000 school districts every year, based on data now collected in the American Community Survey and predictor variables obtained from other sources available at the local area level, such as tax data (U.S. Census Bureau 2009b). A comprehensive treatment of indirect estimation using small area estimation techniques, a methodology that falls outside the scope of this paper, can be found in Rao (2003).

Apart from location sampling, discussed in Section 3.6, this paper also does not address the various methods that have been developed for sampling other types of mini-domains of much interest to social researchers and epidemiologists, domains that are often "hidden populations" in that the activities defining them are clandestine, such as intravenous drug use (Watters and Biernacki 1989). A range of methods has been developed under the assumption that the members of the mini-domains know each other. The broad class of such designs is termed link-tracing designs (see the review by Thompson and Frank 2000). They are adaptive designs in that the units are selected sequentially, with those selected at later stages dependent on those selected earlier (Thompson and Seber 1996; Thompson 2002).

Snowball sampling was one of the early methods of an adaptive, chain-referral sample design. It starts with some initial sample of rare domain members (the seeds), and they in turn identify other members of the domain. While it bears a resemblance to network (multiplicity) sampling (described in Section 3.5), snowball sampling lacks the probability basis of the latter technique, *i.e.*, known, non-zero, selection probabilities for all members of the domain. A version of snowball sampling has been termed respondent-driven sampling (RDS) (Heckathorn 1997, 2007). Volz and Heckathorn (2008) develop a theory for RDS that is based on four assumptions: (1) that respondents know how many members of the network are linked to them (the degree); (2) that respondents recruit others from their personal network at random; (3) that network connections are reciprocal; and

(4) that recruitment follows a Markov process. The need for these modeling assumptions for statistical inference is the difference between chain-referral sample designs and the conventional probability sample designs used in surveys which do not need to invoke such assumptions. It is apparent that RDS is appropriate only for mini-domains for which clear networks exist. The method is used mainly in local area settings, but Katzoff, Sirken and Thompson (2002) and Katzoff (2004) have suggested that the seeds could come from a large-scale survey, such as the U.S. National Health Interview Survey.

This paper focuses on the use of probability sampling methods to produce standard design-based, or direct, estimates for characteristics of rare populations, building on previous reviews (*e.g.*, Kish 1965a; Kalton and Anderson 1986; Kalton 1993a, 2003; Sudman and Kalton 1986; Sudman, Sirken and Cowan 1988; and Flores Cervantes and Kalton 2008). Much of the literature deals with the sampling issues that arise when the rare population is the sole subject of study. However, as noted above, surveys are often required to produce estimates for many different domains as well as for the total population. Section 2 reviews the design issues involved when the survey has design objectives for multiple domains whose members can be identified from the sampling frame. The main part of the paper, Section 3, provides a review of a range of methods that have been used to sample rare populations whose members cannot be identified in advance. The paper ends with some concluding remarks in Section 4.

2. Multi-domain allocations

The issue of sample allocation arises when a survey is being designed to produce estimates for a number of different domains, for subclasses that cut across the domains, as well as for the total population. In most applications, domains vary considerably in size with at least some of them being rare domains.

Assume that there are H mutually exclusive and exhaustive domains that are identified on the sampling frame. Under the commonly made assumptions that the variance of an estimate for domain h can be expressed as V/n_h and that survey costs are the same across domains, the optimum allocation for estimating the overall population mean is $n_h \propto W_h$, where W_h is the proportion of the population in domain h . Assuming that the domain estimates are all to have the same precision, the optimum allocation is $n_h = n/H$ for all domains. These two allocations are in conflict when the W_h vary greatly, as often occurs when the domains are administrative areas of the country, such as states, provinces, counties or districts. In such cases, adopting the optimum allocation for one

objective leads to a serious loss of precision for the other. However, a compromise allocation that falls between the two optimum allocations often works well for both objectives.

Several compromise solutions exist. One, proposed by Kish (1976, 1988), is to determine the domain sample sizes by the following formula:

$$n_h \propto \sqrt{IW_h^2 + (1-I)H^{-2}},$$

where I and $(1-I)$ represent the relative importance of the national estimate and the domain (*e.g.*, administrative district) estimates, respectively. If $I = 1$, the allocation is a proportionate allocation, as optimum for the national estimate, whereas if $I = 0$, the allocation is an equal allocation, as optimum for the domain estimates. The choice of I is highly subjective, but I have found that $I = 0.5$ is often a good starting point, after which a careful review of the allocation can lead to modifications. Bankier (1988) has proposed a similar compromise solution, termed a power allocation. Applied to the current example, the domain sample sizes are determined from $n_h \propto W_h^q$, where q is a power between 0 (equal allocation) and 1 (proportionate allocation). As an example, the 2007 Canadian Community Health Survey was designed to attach about equal importance to the estimates for provinces and health regions. The sample allocation to a province was based on its population size and its number of health regions. Within a province, the sample was allocated between health regions using the Bankier allocation with $q = 0.5$ (Statistics Canada 2008).

A limitation to the Kish and Bankier procedures is that they may not allocate sufficient sample to small domains to produce estimates at the required level of precision. This limitation can be addressed by revising the initial allocations to satisfy precision requirements. An alternative approach addresses this limitation directly: the allocation is determined by fixing a core sample that will satisfy one of the objectives and then supplementing that sample as needed to satisfy the other objective. Singh, Gambino and Mantel (1994) describe such a design for the Canadian Labour Force Survey, with a core sample to provide national and provincial estimates and, where needed, supplemental samples to provide subprovincial estimates of acceptable precision.

The Kish and Bankier schemes assume that the same precision level is required for all small domains. Longford (2006) describes a more general approach in which 'inferential priorities' P_d are assigned to each domain d . As an example, he proposes setting the priorities as $P_d = N_d^a$, where N_d is the population size of domain d and a is a value chosen between 0 and 2. The value $a = 0$ corresponds to the Kish and Bankier equal domain sample size assumption and $a = 2$ corresponds to an overall proportionate

allocation. An intermediate value of a attaches greater priority to larger domains. Longford also extends the approach to incorporate an inferential priority for the overall estimate.

A more general approach to sample allocation is via mathematical programming, as has been proposed by a number of researchers (see, for example, Rodriguez Vera 1982). This approach can accommodate unequal variances across domains, intersecting domains, and multiple estimates for each domain. The U.S. Early Childhood Longitudinal Study – Birth Cohort (ECLS-B) provides an example with intersecting domains, with the sample selected from birth certificate records that contained the requisite domain information. There were 10 domains of interest for the ECLS-B: births classified by race (5 domains), birth weight (3 domains) and twins or non-twins (2 domains). The approach adopted first determined a minimum effective sample size (*i.e.*, the actual sample size divided by the design effect) for each domain. With the 30 cells of the cross-classification of birth weight, race/ethnicity and twin/non-twin treated as strata, an allocation of the sample across the strata was then determined to minimize the overall sample size while satisfying the effective sample size requirements for all the domains (Green 2000).

When there are multiple domains of interest and multi-stage sampling is to be used, a variant of the usual measure of size for probability proportional to size (PPS) sampling can be useful for controlling the sample sizes in the sampled clusters (PSUs, second-stage units, *etc.*), provided that reasonable estimates of the domain population sizes are available by cluster. The requirements that all sampled clusters have approximately the same overall subsample size and that sampled units in each domain have equal probabilities of selection can both be met by sampling the clusters with standard PPS methods, but with a composite measure of size that takes account of the differing sampling rates for different domains (Folsom, Potter and Williams 1987). As an example, in a survey of men in English prisons, the desired sampling fractions were 1 in 2 for civil prisoners (C), 1 in 21 for “star” prisoners who are normally serving their first term of imprisonment (S) and 1 in 45 for recidivists (R). Prisons were selected at the first stage of sampling, with prison i being selected with probability proportional to its composite measure of size $R_i + 2.2S_i + 20.3C_i$, where the multipliers are the sampling rates relative to the rate for recidivists (Morris 1965, pages 303-306).

3. Methods for oversampling rare domains

The main focus of this paper is on the use of probability sampling methods to produce standard design-based, or

direct, estimates for characteristics of rare populations, often minor domains in Kish’s terminology. As preparation for the subsequent discussion, it will be useful to note some features of different types of rare populations that, together with the survey’s mode of data collection, are influential in the choice of sampling methods that can be applied to generate required sample sizes for all domains. Some important features for consideration are summarized below:

- Is a separate frame(s) available for sampling a rare population? Can those sampled be located for data collection? How up-to-date and complete is the frame? If an existing up-to-date frame contains only the rare population (with possibly a few other listings) and provides almost complete coverage, then sampling can follow standard methods. If no single frame gives adequate coverage but there are multiple frames that between them give good coverage, issues of multiple routes of selection arise (Section 3.4).
- Is the rare population concentrated in certain, identifiable parts of the sampling frame, or is it fairly evenly spread throughout the frame? If it is concentrated, disproportionate stratification can be effective (Section 3.2).
- If a sample is selected from a more general population, can a sampled person’s membership in the rare population be determined inexpensively, such as from responses to a few simple questions? If so, standard screening methods may be used (Section 3.1). If accurate determination requires expensive procedures, such as medical examinations, a two-phase design may be useful (Section 3.3). A related issue is whether some members of a rare population consider their membership to be sensitive; the likelihood that members may be tempted to deny their membership may influence the choice of survey administration mode and other aspects of screening.
- Are members of the rare population readily identified by others? If so, some form of network, or multiplicity, sampling may be useful (Section 3.5).
- Are members of the rare population to be found at specific locations or events? If so, location sampling may be useful (Section 3.6).
- Is the rare population defined by a constant characteristic (*e.g.*, race/ethnicity) or by a recent event (*e.g.*, a hospital stay)? The distinction between these two types of characteristics is important in considering the utility of panel surveys for sampling rare populations (Section 3.7).

The following sections review a range of methods for sampling rare populations. Although the methods are discussed individually, some are interrelated and, in practice, a combination of methods is often used.

3.1 Screening

Some form of screening is generally needed when the sampling frame does not contain domain identifiers. This section considers a straightforward application of a screening design in which a large first-phase sample is selected to identify samples of the members of the domains of interest, without recourse to the techniques described in later sections. The first-phase sample size is the minimum sample size that will produce the required (or larger) sample sizes for all of the domains. The minimum first-phase sample size is determined by identifying the required sample size for one of the domains, with all of the sample members of that domain then being included in the second-phase sample. Subsamples of other domains are selected for the second-phase sample at rates that generate the required domain sample sizes. If the survey is designed to collect data for only a subset of the domains (often only one domain), then none of the members of the other domains is selected for the second-phase sample.

Since a very large screening sample size is needed to generate an adequate domain sample size when one (or more) of the domains of interest is a rare population, the cost of screening becomes a major concern. In addition to the sampling methods discussed in later sections, there are several strategies that can be employed to keep costs low:

- Use an inexpensive mode of data collection, such as telephone interviewing or a mail questionnaire, for the screening. The second-phase data collection may be by the same mode or a different mode.
- When possible and useful, permit the collection of screening data from persons other than those sampled. For example, other household members may be able to accurately report the rare population status of the sampled member. See the discussion below and also Section 3.5 on multiplicity sampling.
- When screening is carried out by face-to-face interviewing in a multistage design, it is efficient to select a large sample size in each cluster. Compact clusters can also be used. Costs are reduced, and the precision of domain estimates is not seriously harmed because the average domain sample sizes in the clusters will be relatively small.

One possible means of reducing screening costs is to share the costs across more than one survey. For instance, the child component of the ongoing U.S. National

Immunization Survey (NIS) is a quarterly telephone survey that screens households with landline telephone numbers to locate children aged 19 to 35 months, in order to ascertain vaccination coverage levels (Smith, Battaglia, Huggins, Hoaglin, Roden, Khare, Ezzati-Rice and Wright 2001; U.S. National Center for Health Statistics 2009b). The NIS large-scale screening is also used to identify members of domains of interest for the State and Local Area Integrated Telephone Survey (SLAITS) program, which addresses a variety of other topics over time (U.S. National Center for Health Statistics 2009c). When sharing screening costs across a number of surveys, it is advantageous if the domains for the surveys are fairly disjoint sets in order to minimize the problems associated with screening some respondents into more than one survey.

When no one is at home to complete a face-to-face screening for a household, it may be possible to obtain information from knowledgeable neighbors as to whether the household contains a member of the rare population (e.g., a child under 3 years of age). This approach (which is used in NHANES) can appreciably reduce data collection costs when a large proportion of the households do not contain members of the rare population. However, there is a danger that the approach may result in undercoverage; some protection is provided by requiring that, if the first neighbor interviewed indicates that the household does not include a member of the rare population(s), the other neighbor is also interviewed. Ethical issues also must be considered, particularly for the identification of rare populations that are sensitive in nature.

An extension of the approach of collecting screening information from neighbors is known as focused enumeration. This technique, which is a form of multiplicity sampling (see Section 3.5), involves asking the respondent at each sampled, or “core”, address about the presence of members of the rare population in the n neighboring addresses on either side. In essence, the sample consists of $2n + 1$ addresses for each core address. If the respondent is unable to provide the screening information for one or more of the linked addresses, then the interviewer must make contact at another address. Focused enumeration has been used with $n = 2$ in the British Crime Survey (Bolling, Grant and Sinclair 2008) and the Health Survey of England (Erens, Prior, Korovessis, Calderwood, Brookes and Primatesta 2001) to oversample ethnic minorities. A limitation of the technique is that it will likely produce some (possibly substantial) undercoverage. Evidence of the extent of undercoverage can be obtained by comparing the prevalence of the rare population in the core sample with that in the linked addresses.

In surveys that sample persons by first sampling households, survey designers often prefer to select one person per

household – perhaps allowing two persons to be sampled in large households – to avoid contamination effects and prevent a within-household clustering homogeneity effect on design effects. This design is not always the best (Clark and Steel 2007), and this particularly applies when rare populations are sampled. When rare population members are concentrated in certain households (*e.g.*, minority populations), the size of the screening sample can be appreciably reduced if more than one person – even all eligible persons – can be taken in some households (see Hedges 1973). Elliott, Finch, Klein, Ma, Do, Beckett, Orr and Lurie (2008) suggest that, for oversampling American Indian/Alaskan Native and Chinese minorities in the United States, taking all eligible persons in a household has potential for U.S. health surveys. The NHANES maximizes the number of sampled persons per household. Since each respondent is remunerated for participation, households with more respondents receive more remuneration, a factor thought to increase response rates (Mohadjer and Curtin 2008). Note that within-household homogeneity will have little effect on design effects when the data are analyzed by subgroup characteristics (*e.g.*, age and sex) that cut across households.

The use of large-scale screening to identify rare populations raises three issues, each of which could lead to a failure to achieve planned sample sizes unless precautions are taken. The first results from the fact that, with screening, the sample size for a rare population is a random variable. As a result, the achieved sample size may be larger or smaller than expected. When a minimum sample size is specified for a rare population, it may be wise to determine the sampling fraction to be used to ensure that there is, say, a 90 percent probability that the achieved sample size will be at least as large as the specified minimum. This procedure was used in determining the sampling fractions for the many age, sex and income subdomains for the Continuing Survey of Food Intakes by Individuals 1994-96 (Goldman, Borrud and Berlin 1997).

The second issue raised by large-scale screening is that the overall nonresponse rate must be considered. A sampled member of a rare population will be a nonrespondent if the screener information is not obtained, or if a member of the rare population is identified (perhaps by a proxy informant) but does not respond to the survey items. The overall nonresponse rate may well be much higher than would occur without the screening component. Furthermore, the survey designers must consider the nature of the rare domain and the ways in which members of that domain will react to the survey content. A survey in which new immigrants are asked about their immigration experiences might have a very different response rate than a survey in which war veterans are asked about the medical and other support services they are receiving.

The third issue is that noncoverage can be a significant problem when large-scale screening is used to identify rare populations. One source of noncoverage relates to the sampling frame used for the screener sample. Even though a frame has good overall coverage, its coverage of a rare domain may be inadequate. For example, the noncoverage of a frame of landline telephone numbers is much higher for households of younger people than for the total population. The designers of landline telephone surveys of such rare domains as young children and college students therefore must carefully consider the potential for noncoverage biases. To address the problem of the substantial noncoverage of poor people in telephone surveys, the National Survey of America's Families, which was designed to track the well-being of children and adults in response to welfare reforms, included an area sample of households without telephones in conjunction with the main random digit dialing (RDD) telephone sample (Waksberg, Brick, Shapiro, Flores Cervantes and Bell 1997).

Another source of noncoverage is a failure to identify some members of the rare population at the screening stage. In particular, when a survey aims to collect data only for members of a rare domain, some screening phase respondents may falsely report, and some interviewers may falsely record, that the sampled persons are not members of that domain. These misclassifications may be inadvertent or they may be deliberately aimed at avoiding the second-phase data collection. Misclassification error can give rise to serious levels of noncoverage, particularly when the rare population classification is based on responses to several questions, misreports to any one of which leads to a misclassification (Sudman 1972, 1976). When the survey oversamples one or more rare domains as part of a survey of the general population, misclassifications are uncovered at the second phase, thus avoiding noncoverage. However, misclassifications still result in a smaller sample sizes for rare domains; in addition, the variation in sampling weights between respondents selected as members of the rare domain and those sampled as members of another domain can lead to a serious loss of precision. Noncoverage is more likely to arise when screener data are collected from proxy informants. It is a particular problem with focused enumeration.

In a number of surveys of rare populations, the proportion of rare population members identified has been much lower than prevalence benchmarks. For example, the 1994 NIS had an appreciable shortfall in the identified proportion of children aged 19 to 35 months (4.1 percent compared to the predicted rate of 5 percent) (Camburn and Wright 1996). In the National Longitudinal Survey of Youth of 1997, only 75 percent of youth aged 12 to 23 years were located (Horrigan, Moore, Pedlow and Wolter 1999). These findings could be the result of higher nonresponse rates for

members of the rare population, frame noncoverage of various types, or misclassifications of domain membership. To produce the required sample size, an allowance for under-representation must be made at the design stage. The noncoverage of an age domain appears to be greatest at the domain boundaries, perhaps because respondents do not know exact ages (with those falsely screened out being lost and those falsely screened in being detected and dropped later) or because of deliberate misreporting to avoid the follow-up interview. To counteract this effect, it can be useful to start with an initial screening for all household members or for a broader age range and then narrow down to the required age range later on.

Weighting adjustments can be used in an attempt to mitigate biases caused by nonresponse and noncoverage, but they are necessarily imperfect. Adjustments for a domain specific level of nonresponse require knowledge of the domain membership of nonrespondents, but that is often not available. Adjustments for noncoverage of a rare domain require accurate external data for the domain, data that are often not available. Indeed, one of the purposes for some rare domain surveys is to estimate the domain size. Noncoverage is a major potential source of error in the estimation of domain size.

3.2 Disproportionate stratification

A natural extension of the screening approach is to try to identify strata where the screening will be more productive. In the ideal circumstance, one or more strata that cover all of the rare population and none from outside that population are identified. That case requires no screening process. Otherwise, it is necessary to select samples from all the strata (apart from those known to contain no rare population members) to have complete coverage of the rare population. The use of disproportionate stratification, with higher sampling fractions in the strata where the prevalence of the rare population is higher, can reduce the amount of screening needed.

3.2.1 Theoretical background

Consider initially a survey designed to provide estimates for a single rare population. Waksberg (1973) carried out an early theoretical assessment of the value of disproportionate stratification for this case. Subsequent papers on this topic include those by Kalton and Anderson (1986) and Kalton (1993a, 2003). The theoretical results show that three main factors must be considered in determining the effectiveness of disproportionate stratification for sampling a single rare population: the prevalence rate in each stratum, the proportion of the rare population in each stratum, and the ratio of the full cost of data collection for members of the rare population to the screening cost involved in identifying

members of that population. If it is assumed that (1) the element variances for the rare population are the same across strata and (2) the costs of data collection for members of the rare and non-rare populations are the same across strata, then, with simple random sampling within strata, the optimum sampling fraction in stratum h for minimizing the variance of an estimated mean for the rare population, subject to a fixed total budget, is given by

$$f_h \propto \sqrt{\frac{P_h}{P_h(c-1) + 1}},$$

where P_h is the proportion of the units in stratum h that are members of the rare population and c is the ratio of the data collection cost for a sampled member of the rare population to the cost for a member of the non-rare population (Kalton 1993a). The following formula provides the ratio of the variance of the sample mean with the optimum disproportionate stratified sampling fractions to that with a proportionate stratified sample of the same total cost:

$$R = \frac{[\sum A_h \sqrt{P(c-1) + P/P_h}]^2}{P(c-1) + 1},$$

where A_h is the proportion of the rare population in stratum h and P is the prevalence of the rare population in the full population.

In general, the variability in the optimum sampling fractions across the strata, and the gains in precision for the sample mean, decline as c increases. Thus, if the main survey data collection cost is high – as, for instance, when the survey involves an expensive medical examination – or if the screening cost is very low, then disproportionate stratification may yield only minor gains in precision.

When the main data collection cost adds nothing to the screening cost, the ratio of main data collection cost to screening cost will be $c = 1$. In this limiting situation, the formulas given above simplify to $f_h \propto \sqrt{P_h}$ and $R = (\sum \sqrt{A_h W_h})^2$, where W_h is the proportion of the total population in stratum h . These simple formulas provide a useful indication of the maximum variation in optimum sampling fractions and the maximum gains in precision that can be achieved. The square root function in the optimum sampling fraction formula makes clear that the prevalences in the strata must vary a good deal if the sampling fractions are to differ appreciably from a proportionate allocation. For example, even if the prevalence in stratum A is four times as large as that in stratum B, the optimum sampling fraction in stratum A is only twice as large as that in stratum B. The gains in precision ($1 - R$) are large when A_h is large when W_h is small and vice versa. With only two strata, a stratum with a prevalence five times as large as the overall prevalence (*i.e.*, $P_h/P = 5$) will yield gains in precision of 25 percent or more ($(1 - R) \geq 0.25$) only if that stratum

includes at least 60 percent of the rare population (Kalton 2003, Table 1).

In summary, while generally useful, disproportionate stratification will yield substantial gains in efficiency only if three conditions hold: (1) the rare population must be much more prevalent in the oversampled strata; (2) the oversampled strata must contain a high proportion of the rare population; and (3) the cost of the main data collection per sampled unit must not be high. In many cases, not all of these three conditions can be met, in which case the gains will be modest.

Furthermore, the results presented above are based on the assumption that the true prevalence of the rare population in each stratum is known, whereas in practice it will be out of date (for example, based on the last census) or will perhaps simply have been guesstimated. Errors in the prevalence estimates will reduce the precision gains achieved with disproportionate stratification and could even result in a loss of precision. A major overestimation of the prevalence of the rare population, and hence of the optimum sampling fraction, in the high-density stratum can result in a serious loss of precision for the survey estimates. It is therefore often preferable to adopt a conservative strategy, that is, to adopt a somewhat less disproportionate allocation, one that moves in the direction of a proportionate allocation.

3.2.2 Applications

When area sampling is used, data available from the last census and other sources can be used to allocate the area clusters to strata based on their prevalence estimates for the rare population. See Waksberg, Judkins and Massey (1997) for a detailed investigation of this approach for oversampling various racial/ethnic populations and the low-income population using U.S. census blocks and block groups as clusters. Based on data from the 1990 Census, Waksberg and his colleagues found that the approach generally worked well for Blacks and Hispanics but not for the low-income population. While the low-income population did exhibit high concentrations in some blocks and block groups, those areas did not cover a high proportion of that population.

When the survey designers have access to a list frame with names, the names can be used to construct strata of likely members of some racial/ethnic groups. This situation arises, for instance, with lists of names and telephone numbers and when names are merged onto U.S. Postal Service (USPS) Delivery Sequence File addresses (no name merge is made in some cases). The allocation to strata can be based on surnames only or on a combination of surname and first name (and even other names also). Since women often adopt their husbands' surnames, the allocation is generally more effective for men than women. Names can

be reasonably effective for identifying Hispanics, Filipinos, Vietnamese, Japanese and Chinese, but not Blacks. A number of lists of names associated with different racial/ethnic groups have been compiled, such as the list of Spanish names compiled by the U.S. Census Bureau for the 1990s (Word and Perkins 1996). Several commercial vendors have developed complex algorithms to perform racial/ethnic classifications based on names (see Fiscella and Fremont 2006 for further details). The use of names in identifying race and ethnicity has been of considerable interest to epidemiologists and demographers, who have conducted a number of evaluations of this method (*e.g.*, Lauderdale and Kestenbaum 2000; Elliott, Morrison, Fremont, McCaffrey, Pantoja and Lurie 2009). They often assess the effectiveness of the method in terms of positive predictive value and sensitivity, which are the equivalents of prevalence and the proportion of members of the domain who are identified as such by the instrument used for the classification. In the sampling context, besides limitations in the instrument, researchers also need to take into account that sometimes names are not available and that some available names may be incorrect (for example, with address-based sampling, the names may be out-of-date, because the original family has moved out and a new family has moved into an address). These additional considerations serve to reduce the effectiveness of the name stratification, and depending on the particular circumstances, the reduction in effectiveness may be sizable.

As with stratification in general, the stratification factors used for sampling rare populations do not have to be restricted to objective measures. They can equally be subjective classifications. The only consideration is how well they serve the needs of the stratification (see Kish 1965b, pages 412-415, for an example of the effectiveness of the use of listers' rapid classifications of dwellings into low, medium or high socio-economic status for disproportionate stratification). Elliott, McCaffrey, Perlman, Marshall and Hambarsoomians (2009) describe an effective application of subjective stratification for sampling Cambodian immigrants in Long Beach, California. A local community expert rated all individual residences in sampled blocks as likely or unlikely to contain Cambodian households, based on externally observable cultural characteristics such as footwear outside the door and Buddhist altars. The residences allocated to the "likely" stratum (approximately 20 percent) were then sampled at four times the rate than the rest.

Sometimes, when the survey is concerned with producing estimates only for a very rare population, disproportionate stratification may still require an excessive amount of screening. In that circumstance, it may be necessary to sample from the strata where the prevalence is

highest, dropping the other strata and accepting some degree of noncoverage (or redefining the survey population to comprise only members of the rare population in the strata that were sampled). The Hispanic Health and Nutrition Examination Survey of 1982-84 (HHANES) provides an illustration. For its samples of Mexican Americans in the Southwest and Puerto Ricans in the New York City area, the HHANES sampled only from counties with large numbers and/or percentages of Hispanics, based on 1980 Census counts (Gonzalez, Ezzati, White, Massey, Lago and Waksberg 1985).

As another example of this approach, Hedges (1979) describes a procedure for sampling a minority population that is more concentrated in some geographical districts, such as census enumeration districts. In this procedure, the districts are listed in order of their prevalence of members of the rare population (obtained, say, from the last census), and then the survey designers produce Lorenz curves of the cumulative distribution of rare population prevalence and the cumulative distribution of the proportions of rare population members covered. With the cumulative prevalence declining as the cumulative coverage increases, the survey designers can use these distributions to select the combination of prevalence and proportion covered that best fulfills their requirements. The issue then to be faced is whether to make inferences to the covered population, or whether to make inferences to the full population by applying population weighting adjustments in an attempt to address the noncoverage bias.

When a domain is very rare but a portion of it is heavily concentrated in a stratum, researchers sometimes sample that stratum at a rate much higher than the optimum in order to generate a sizable number of cases. Although this approach may produce a large sample of the rare population, the effective sample size (*i.e.*, the sample size divided by the design effect) will be smaller than if the optimum sampling fractions had been used. Thus, from the perspective of the standard survey design-based mode of inference, this approach is not appropriate. However, the researchers using this approach often argue for a model-based mode of inference in which the sampling weights are ignored. In my view, ignoring the sampling weights is problematic. However, discussion of this issue is outside the scope of this paper.

3.3 Two-phase sampling

The screening approach treated in Sections 3.1 and 3.2 assumes that identification of rare population members is relatively easy. When accurate identification is expensive, a two-phase design can be useful, starting with an imperfect screening classification at the first phase, to be followed up with accurate identification for a disproportionate stratified

subsample at the second phase. Whether the two-phase approach is cost-effective depends in part on the relative costs of the imperfect classification and accurate identification: since the imperfect classifications use up some of the study's resources, they must be much less expensive than the accurate identification. Deming (1977) suggests that the ratio of the per-unit costs of the second- to the first-phase data collections should be at least 6:1. Also, the imperfect classification must be reasonably effective in order to gain major benefits from a second-phase disproportionate stratification.

Two- or even three-phase sampling can often be useful in medical surveys of persons with specific health conditions. The first phase of the survey often consists of a screening questionnaire administered by survey interviewers, and the second phase is generally conducted by clinicians, often in a medical center. As one example, in a survey of epilepsy in Copiah County, Mississippi, Haerer, Anderson and Schoenberg (1986) first had survey interviewers administer to all households in the county a questionnaire that had been pretested to ensure that it had a high level of sensitivity for detecting persons with epilepsy. To avoid false negatives at this first phase, a broad screening net was used in identifying persons who would continue to the second phase. All those so identified were the subjects for the second phase of the survey, which consisted of brief neurological examinations conducted by a team of four senior neurologists in a public health clinic.

A second example illustrates the use of another survey to serve as the first-phase data collection for studying a rare domain. In this case, the Health and Retirement Study (HRS) was used as the first phase for a study of dementia and other cognitive impairment in adults aged 70 or older. The HRS collects a wide range of measures on sample respondents, including a battery of cognitive measures. Using these measures, the HRS respondents were allocated to five cognitive strata, with a disproportionate stratified sample being selected for the second phase. The expensive second-phase data collection consisted of a 3- to 4-hour structured in-home assessment by a nurse and neuropsychology technician. The results of the assessment were then evaluated by a geropsychiatrist, a neurologist and a cognitive neuroscientist to assign a preliminary diagnosis for cognitive status, which was then reassessed in the light of data in the person's medical records (Langa, Plassman, Wallace, *et al.* 2005).

A third example is a three-phase design that was used in a pilot study to identify persons who would qualify for disability benefits from the U.S. Social Security Administration if they were to apply for them (Maffeo, Frey and Kalton 2000). At the first phase, a knowledgeable household respondent was asked to provide information about the

disability beneficiary status and impairment status of all adults aged 18 to 69 years in the household. At the second phase, all those classified into a stratum of severely disabled nonbeneficiaries and samples of the other strata were interviewed in person and were then reclassified as necessary into likely disability strata for the third phase. At the third phase, a disproportionate stratified sample of persons was selected to undergo medical examinations in mobile examination centers.

A fairly common practice with two-phase designs is to take no second- (or third-) phase sample from the stratum of those classified as nonmembers of the rare domain based on their responses at the previous phase. The proportion of the population in that stratum is usually very high, and the prevalence of the rare domain in it is very low (indeed, as in the Haerer, Anderson and Schoenberg (1986) study, the stratum is often conservatively defined with the aim of avoiding the inclusion of those who might possibly be members of the rare domain). As a result, a moderate-sized sample from this stratum will yield almost no members of the rare domain. However, the cut-off strategy of taking no sample from this stratum is risky. If the prevalence of the rare domain in this large stratum is more than minimal, a substantial proportion of the domain may go unrepresented in the final sample.

3.4 Multiple frames

Sometimes sampling frames exist that are more targeted on a rare population than a general frame, but they cover only part of the rare population. In this situation, it can be efficient to select the sample from more than one frame. For example, in the common case of oversampling ethnic minorities, there is sometimes a list frame available. The persons on the list can be classified based on their names as being likely to belong to a given ethnic group (*e.g.*, Chinese, Korean, Pacific Islanders, Vietnamese) to create a second, incomplete sampling frame from which to sample, in addition to a more complete frame that has a lower prevalence of the rare population (see, *e.g.*, Elliott *et al.* 2008; Flores Cervantes and Kalton 2008). As with disproportionate stratification (Section 3.2), major benefits derive from this approach only when the second frame has a high prevalence and covers a sizable fraction of the rare population. See Lohr (2009) for a review of the issues involved in sampling from multiple frames.

With multiple frames, some members of the rare population may be included on several frames, in which case they may have multiple routes of being selected into the sample. There are three broad approaches for addressing these multiplicities (Anderson and Kalton 1990; Kalton and Anderson 1986). When all the frames are list frames, as sometimes occurs in health studies, it may be possible to

combine the frames into a single unduplicated list; however, this can often involve difficult record linkage problems. An alternative approach is to make the frames non-overlapping by using a unique identification rule that associates each member of the rare population with only one of the frames, treating the listings on the other frames as blanks (Kish 1965b, pages 388-390). Samples are selected from each of the frames without regard to the duplication, but only the non-blank sampled listings are accepted for the final sample. This approach works best when searches can be made for each sampled unit on the other frames; if the frames are put in a priority order and the unit is found on a prior frame to the one from which the selection was made, the sampled listing would be treated as a blank. In this case, the frames are strata; the sampled units are treated as subclasses within the strata, allowing for the blank listings (Kish 1965b, pages 132-139), and the analysis follows standard methods.

The use of the unique identification approach can, however, be inefficient when the persons sampled from one frame have to be contacted to establish whether their listings are to be treated as real or blank for that frame. In this case, it is generally more economical to collect the survey data for all sampled persons (*i.e.*, to accept the multiple routes of selection). There are, however, exceptions, as in the case of the National Survey of America's Families. That survey used a combination of an area frame and an RDD telephone frame, with the area frame being used to cover only households without telephones (Waksberg, Brick *et al.* 1997). It proved to be efficient to conduct a quick screening exercise with households on the area frame to eliminate households with telephones, retaining only the non-telephone households for the survey.

There are two general approaches for taking multiple routes of selection into account in computing selection probabilities (Bankier 1986; Kalton and Anderson 1986). One method calculates each sampled unit's overall selection probability across all the frames and uses the inverse of that probability as the base weight for the analysis (leading to the Horvitz-Thompson estimator). For example, the overall selection probability for sampled unit i on two frames is $p_i = (p_{1i} + p_{2i} - p_{1i}p_{2i}) = [1 - (1 - p_{1i})(1 - p_{2i})]$, where p_{fi} is the probability of the unit's selection from frame $f = 1, 2$. A variant is to replace the overall selection probability with the expected number of selections (leading to the Hansen-Hurwitz estimator), which is easier to compute when multiple frames are involved. With only two frames, the expected number of selections is $(p_{1i} + p_{2i})$. When selection probabilities are small, there is little difference between these two estimators.

Adjustments to compensate for nonresponse and to calibrate sample totals to known population totals can either be made to the overall selection probabilities p_i or they can

be made to the p_{fi} individually. A problem that can occur is that the survey designers do not know whether a nonresponding unit sampled from one frame is on another frame since that information is only collected in the interview. In this situation the p_i for nonresponding units cannot be directly computed and must be estimated in some fashion. When adjustments are made to the p_{fi} individually, it is not possible to form nonresponse weighting classes that take membership on other frames into account. Instead, the designers must assume that, within weighting classes, the response rates are the same no matter how many frames a unit is on.

In general, the application of the approach described above requires knowledge of each sampled unit's selection probabilities for all of the frames, information that is not always available. When selection probabilities are not known for frames other than the frame(s) from which the unit is sampled (but presence/absence on the frames is known), an alternative approach, termed a weight share method by Lavallée (1995, 2007), can be used. Unbiased estimates of population totals are obtained if the weight for unit i is given by $w_i = \sum_j \alpha_{ij} w'_{ij}$ where α_{ij} are any set of constants such that $\sum_j \alpha_{ij} = 1$ when summed across the j frames, $w'_{ij} = 1/p_{ij}$ if unit i is selected from frame j with probability p_{ij} and $w'_{ij} = 0$ otherwise (Kalton and Brick 1995; Lavallée 2007). For many applications, it is reasonable to set $\alpha_{ij} = \alpha_j$ and then a good choice of α_j is $\alpha_j = \tilde{n}_j / \sum \tilde{n}_j$, where \tilde{n}_j is the effective sample size based on some average design effect (Chu, Brick and Kalton 1999).

The second general approach for dealing with multiple routes of selection uses the multiple-frame methodology introduced by Hartley (1974), and the subject of much recent research (see, e.g., Lohr and Rao 2000 and 2006 and the references cited in those papers). In the case of two frames (A and B), the population can be divided into three mutually exclusive subsets labeled $a = A \cap \bar{B}$, $b = \bar{A} \cap B$ and $ab = A \cap B$. The sample can be divided into samples from a , b and ab , where the ab sample can be separated into respondents sampled from frame A and those sampled from frame B . The samples in subsets a and b have only one route of selection, and hence are readily handled in estimation. Totals for ab could be estimated from the sample from frame A or the sample from frame B , say, \hat{Y}_{ab}^A or \hat{Y}_{ab}^B . The Hartley methodology takes a weighted average of these two estimators, $\hat{Y}_{ab} = \theta \hat{Y}_{ab}^A + (1 - \theta) \hat{Y}_{ab}^B$, where θ is chosen to minimize the variance of \hat{Y}_{ab} , taking into account that sample sizes and design effects differ between the two samples. Note that the dual-frame methodology is estimator specific, with different values of θ for different estimators. Skinner (1991), Skinner and Rao (1996) and Lohr and Rao (2006) have proposed an alternative, pseudo-maximum likelihood estimation approach that has the

attraction of avoiding the problems associated with different values of θ for different variables. Wu and Rao (2009) propose a multiplicity-based pseudo empirical likelihood approach for multiple frame surveys, including what they term a single-frame multiplicity-based approach that incorporates Lavallée's weight share method as described above.

When a dual- or multiple-frame design is used, it is often the case that one frame has complete coverage but a low prevalence of the rare population (e.g., an area frame) and the other frame(s) has a much higher prevalence of the rare population but incomplete coverage. Metcalf and Scott (2009), for example, combined an area sample with an electoral roll sample for the Auckland Diabetes, Heart and Health Survey, in which Pacific Islanders, Maoris and older people were domains of special interest. The electoral roll frame had the advantage of containing information about electors' ages, as well as a special roll on which those who considered themselves to be of Maori descent could enroll. Furthermore, many people of Pacific descent could likely be identified by their names, since Pacific languages use fewer letters than English. A disproportionate stratified sample was selected from the electoral roll frame to oversample the domains of interest, and the sample from the area frame brought in people not on the electoral rolls.

The National Incidence Study of Child Abuse and Neglect provides an example of a more complex situation (Winglee, Park, Rust, Liu and Shapiro 2007). That survey used many frames to increase its overall coverage of abused and neglected children. Child Protective Services (CPS) agencies in the sampled PSUs were the basis of the main sampling frame, while police, hospitals, schools, shelters, daycare centers and other agencies were the sources of other frames. The samples from CPS agencies were selected from list frames, but the samples from other agencies were drawn by sampling agencies, constructing rosters of relevant professional staff, and sampling staff who acted as informants about maltreated children. With these procedures, duplication across agencies cannot be ascertained, except in the case of CPS agencies and any of the other agencies. The design was therefore treated as a dual-frame design, with CPS as one frame and the combination of the other frames as the second frame (i.e., assuming no overlap between the other frames).

3.5 Network sampling

Network (or multiplicity) sampling expands on the standard screening approach by asking sampled persons (or addresses) to also serve as proxy informants to provide the screening information for persons who are linked to them in a clearly specified way (Sudman *et al.* 1988; Sirken 2004, 2005). Relatives such as parents, siblings and children are

often used as the basis of linkages. A key requirement is that every member of the linkage must know and be willing to report the rare population membership statuses of all those linked to them. In a pilot study of male Vietnam veterans, Rothbart, Fine and Sudman (1982) included aunts and uncles as informants as well as parents and siblings, but found that aunts and uncles identified far fewer Vietnam veterans than expected. This apparent failure of aunts and uncles to report some veterans gives rise to a potential sampling bias, thus making their inclusion in the linkage rules problematic.

The multiple routes of selection with network sampling need to be taken into account in determining selection probabilities in a similar manner to that described for multiple frames in the previous section. Conceptually, one can consider each member of the rare population divided into, say, l parts corresponding to the l informants for that member; it is then these parts that are sampled for the survey. See Lavallée (2007) for some theory behind the technique.

When network sampling is used in surveys that collect data on the characteristics of rare population members, direct contact must be made with the members of the rare population identified by the initial informant. In this case, the informant has to be able to provide contact information for the rare population members. The linkage definition may be structured to facilitate the follow-up data collection. For example, with face-to-face interviewing, the linkage may be restricted to relatives living in a defined area close to the informant.

Sudman and Freeman (1988) describe the application of network sampling in a telephone survey about access to health care, in which an oversample of persons with a chronic or serious illness was required. During an initial contact with the head of the household, linkages to the respondent's or spouse's parents, stepparents, siblings, grandparents and grandchildren under age 18 were identified and data were collected on their health status. The use of this network sampling design increased the number of chronically or seriously ill adults identified by about one-third. However, about one in eight of the initial network informants with relatives were unable or unwilling to provide illness information for their network members, and 70 percent did not provide complete location information, including 28 percent who provided neither name nor location information (thus making tracing impossible). The use of network sampling led to some false positives (persons reported as being chronically or seriously ill by the initial respondent but reporting themselves as well). A more serious concern is that the survey was not able to provide information on false negatives (this would have required following up a sample of network members reported to be well by the initial informant).

Some forms of linkage have the added benefit that they can incorporate some rare population members who are not on the original sampling frame and would therefore otherwise be a component of noncoverage. For example, Brick (1990) describes a field test for the telephone-based National Household Education Survey (NHES) that used multiplicity sampling to increase the sample of 14- to 21-year-olds, with a focus on school drop-outs. In a subsample of households, all women aged 28 to 65 were asked to provide information for all their 14- to 21-year-old children currently living elsewhere. Some of these children lived in telephone households and hence had two routes of selection. Others lived in non-telephone households and hence would not have been covered by the survey; their inclusion via the multiplicity design increased the coverage rate in 1989 by about 5 percent. However, the response rate for out-of-household youth was much lower than that for in-household youth because of failure to reach the youth, particularly the youth living in non-telephone households.

Tortora, Groves and Peytcheva (2008) provide another example, in this case using multiplicity sampling in an attempt to cover persons with only mobile telephones via an RDD sample of landline telephone numbers. Respondents to the RDD survey (itself a panel survey) were asked to provide information about parents, siblings and adult children living in mobile-only households. The results demonstrate some of the general issues with multiplicity sampling: knowledge about the mobile-only status of the network members depended on the cohesion of the network; there was widespread unwillingness to provide mobile telephone numbers; and many of those identified as mobile-only households in fact also had a landline telephone.

Network sampling has not been widely used in practice for surveys of rare population members. Some of the limitations of the method are illustrated by the studies described above. There is the risk that the sampled informant may not accurately report the rare population status of other members of the linkage, either deliberately or through lack of knowledge. Nonresponse for the main survey data collection is another concern. In addition, ethical issues can arise when sampled persons are asked about the rare population membership of those in their linkage when that membership is a sensitive matter. The benefits of network sampling are partially offset by the increased sampling errors arising from the variable weights that the method entails, and by the costs of locating the linked rare population members.

3.6 Location sampling

Location sampling is widely used to sample populations that have no fixed abode for both censuses and surveys: nomads may, for example, be sampled at waterpoints when

they take their animals for water, and homeless persons may be sampled at soup kitchens when they go for food (*e.g.*, Kalton 1993a; Ardilly and Le Blanc 2001). A central feature of such uses of location sampling is that there is a time period involved, resulting in issues of multiplicity (Kalsbeek 2003). A serious concern with the use of the technique is that it fails to cover those who do not visit any of the specified locations in the particular time period.

Location sampling is used to sample rare mobile populations such as passengers at airports and visitors to a museum or national park. In such cases, the question arises as to whether the unit of analysis should be the visit or the visitor. When the visit is the appropriate unit, no issues of multiplicity arise (see, for example, the report on the U.S. National Hospital Discharge Survey by DeFrances, Lucas, Buie and Golosinskiy 2008). However, when the visitor is the unit of analysis, the fact that visitors may make multiple visits during the given time period must be taken into account (Kalton 1991; Sudman and Kalton 1986). One approach is to treat visits as eligible only if they are the first visits made during the time period for the survey. Another approach is to make multiplicity adjustments to the weights in the analysis; however, determining the number of visits made is problematic because some visits will occur after the sampled visit.

Location sampling has also been used for sampling a variety of rare – often very rare – populations that tend to congregate in certain places. For example, Kanouse, Berry and Duan (1999) employed the technique to sample street prostitutes in Los Angeles County by sampling locations where street prostitution was known to occur, and by sampling time periods (days and shifts within days). Location (center) sampling has also been used to sample legal and illegal immigrants in Italy (Meccati 2004). For a 2002 survey of the immigrant population of Milan, 13 types of centers were identified, ranging from centers that provide partial lists from administrative sources (*e.g.*, legal and work centers, language courses), centers that have counts of those attending (*e.g.*, welfare service centers, cultural associations), to centers with no frame information (*e.g.*, malls, ethnic shops).

Location sampling has often been used to sample men who have sex with men, with the locations being venues that such men frequent, such as gay bars, bathhouses and bookstores (Kalton 1993b, MacKellar, Valleroy, Karon, Lemp and Janssen 1996). Based on a cross-sectional telephone survey, Xia, Tholandi, Osmond, Pollack, Zhou, Ruiz and Catania (2006) found that men who visited gay venues more frequently had higher rates of high-risk sexual behaviors and also that the rates of high-risk behaviors varied by venue. These findings draw attention to the difficulty of generating a representative sample by location sampling.

McKenzie and Mistiaen (2009) carried out an experiment to compare location (intercept) sampling with both area sampling and snowball techniques, for sampling Brazilians of Japanese descent (Nikkei) in Sao Paulo and Parana. The locations included places where the Nikkei often went (*e.g.*, a sports club, a metro station, grocery stores and a Japanese cultural club) and events (*e.g.*, a Japanese film and a Japanese food festival). Based on this experiment, they conclude that location sampling (and snowball sampling) oversampled persons more closely connected with the Nikkei community and thus did not produce representative samples. This not-unexpected finding highlights the concern about the use of location sampling for sampling rare populations in general, although not for sampling visits to specified sites.

3.7 Accumulating or retaining samples over time

When survey data collection is repeated over time, survey designers can take advantage of that feature in sampling rare populations (Kish 1999). An important distinction to be made is that between repeated and panel surveys. Samples of rare population members can readily be accumulated over time in repeated surveys. For example, the U.S. National Health Interview Survey is conducted on a weekly basis with nationally representative samples; samples of rare populations can be accumulated over one or more years until a sufficient sample size is achieved (U.S. National Center for Health Statistics 2009a). With accumulation over time, the estimates produced are period, rather than point-in-time, estimates that can be difficult to interpret when the characteristics of analytic interest vary markedly over time (Citro and Kalton 2007). For example, how is a 3-year period poverty rate for a rare minority population to be interpreted when the poverty rate has varied a great deal over the period?

In considering the sampling of rare populations in panel surveys, it is important to distinguish between rare populations that are defined by static versus non-static characteristics. No accumulation over time can be achieved in panel surveys for rare populations defined by static characteristics such as race/ethnicity. However, if a sample of a static rare population is taken at one point in time, it can be useful to follow that sample in a panel to study that population's characteristics at later time points, possibly with supplementary samples added to represent those who entered that population after the original sample was selected. Fecso, Baskin, Chu, Gray, Kalton and Phelps (2007) describe how this approach has been applied in sampling U.S. scientists and engineers over a decade. For the decade of the 1990s, the National Survey of College Graduates (NSCG) was conducted in 1993 with a stratified sample of college graduates selected from the 1990 Census of Population long-form sample records. Those found to be

scientists or engineers were then resurveyed in the NSCG in 1995, 1997 and 1999. To represent new entrants to the target population, another survey – the Survey of Recent College Graduates – was conducted in the same years as the NSCG. A subsample of the recent college graduates was added in to the next round of the NSCG panel on each occasion.

Panel surveys can be used to accumulate samples of non-static rare populations, especially persons experiencing an event such as a birth or a divorce. The U.S. National Children's Study, for instance, plans to follow a large sample of eligible women of child-bearing age over a period of about four years, enrolling those who become pregnant in the main study, a longitudinal study that will follow the children through to age 21 (National Children's Study 2007, Michael and O'Muircheartaigh 2008).

Finally, a large sample can be recruited into a panel and provide data that will identify members of a variety of rare populations that may be of future interest. They are then followed in the panel and, based on their rare population memberships, included in the samples for the surveys for which they qualify. Körner and Nimmergut (2004) describe a German "access panel" that could be used in this way, and there are now several probability-based Web panels that can serve this purpose (Callegaro and DiSogra 2008). However, a serious concern with such panels is the low response rates that are generally achieved.

4. Concluding remarks

This paper has presented a brief overview of the range of methods used in sample surveys for sampling and oversampling rare populations, primarily those classified by Kish as minor domains (the references cited provide more details). Although the methods have been discussed separately, in practice they are often combined, particularly when there are several rare domains of interest. As an example, the California Health Interview Survey, conducted by telephone, has used a combination of disproportionate stratification (oversampling telephone exchanges where the prevalence of the Korean and Vietnamese populations of interest is higher) and a dual-frame design (RDD methods supplemented with a frame of likely Korean and Vietnamese names). In many cases, the art of constructing an effective probability sample design for a rare population is to apply some combination of methods in a creative fashion.

As another example, the Pew Research Center telephone survey of Muslim Americans employed three sampling methods to sample this very rare population (Pew Research Center 2007). One component of the design was a geographically stratified RDD sample, with disproportionate stratified

sampling from strata defined in terms of the prevalence of Muslim Americans. The stratum with the lowest prevalence was treated as a cut-off stratum and excluded. The second component was a recontact sample of Muslim Americans drawn from Pew's interview database of recent surveys. The third component was an RDD sample selected from a list of likely Muslim Americans provided by a commercial vendor. To avoid duplicate routes of selection between the geographical strata and the commercial vendor list, telephone numbers selected from the geographical strata were matched against the commercial vendor list and dropped from the geographical strata sample if a match was found.

Not only are the various sampling techniques often used in combination in sample designs for rare populations, but several of the techniques are interrelated. For example, multiple frames can be treated by unique identification (see Section 3.4), which in effect is simply disproportionate stratification. Whereas the whole population is classified into strata for disproportionate stratification, the same approach is adopted with two-phase sampling, but the classification into strata is applied only to members of the first-phase sample. The theory of network sampling is similar to that of multiple-frame sampling, when the latter technique uses inverse overall selection probabilities as weights in the analysis. These interrelationships help to explain the similarities in the theoretical underpinnings of the techniques.

Acknowledgements

I would like to thank Daniel Levine and Leyla Mohadjer for helpful reviews of a draft of this paper, Daifeng Han and Amy Lin for constructive comments on an earlier, shorter version of the paper, and to Mike Brick, Marc Elliott, and Jon Rao for advice on some specific points.

References

- Anderson, D.W., and Kalton, G. (1990). Case-finding strategies for studying rare chronic diseases. *Statistica Applicata*, 2, 309-321.
- Ardilly, P., and Le Blanc, D. (2001). Sampling and weighting a survey of homeless persons: A French example. *Survey Methodology*, 27, 109-118.
- Bankier, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *American Statistician*, 42, 174-177.
- Bolling, K., Grant, C. and Sinclair, P. (2008). *2006-07 British Crime Survey (England and Wales)*. Technical Report. Volume I. Available at <http://www.homeoffice.gov.uk/rds/pdfs07/bcs0607tech1.pdf>.

- Brick, J.M. (1990). Multiplicity sampling in an RDD telephone survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 296-301.
- Callegaro, M., and DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, 72, 1008-1032.
- Camburn, D.P., and Wright, R.A. (1996). *Predicting eligibility rates for rare populations in RDD screening surveys*. Available at http://www.cdc.gov/nis/pdfs/sample_design/camburn1996.pdf.
- Chu, A., Brick, J.M. and Kalton, G. (1999). Weights for combining surveys across time or space. *Bulletin of the International Statistical Institute, Contributed Papers*, 2, 103-104.
- Citro, C.F., and Kalton, G. (Eds.) (2007). *Using the American Community Survey: Benefits and Challenges*. Washington, DC: National Academies Press.
- Clark, R.G., and Steel, D.G. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society*, 170, Series A, 63-82.
- DeFrances, C.J., Lucas, C.A., Buie, V.C. and Golosinskiy, A. (2008). *2006 National Hospital Discharge Survey*. National Health Statistics Reports Number 5. U.S. National Center for Health Statistics, Hyattsville, MD.
- Deming, W.E. (1977). An essay on screening, or on two-phase sampling, applied to surveys of a community. *International Statistical Review*, 45, 29-37.
- Durr, J.-M. (2005). The French new rolling census. *Statistical Journal of the United Nations Economic Commission for Europe*, 22, 3-12.
- Elliott, M.N., Finch, B.K., Klein, D., Ma, S., Do, D.P., Beckett, M.K., Orr, N. and Lurie, N. (2008). Sample designs for measuring the health of small racial/ethnic subgroups. *Statistics in Medicine*, 27, 4016-4029.
- Elliott, M.N., Morrison, P.A., Fremont, A., McCaffrey, D.F., Pantoja, P. and Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services Outcomes Research Methods*, 9, 69-83.
- Elliott, M.N., McCaffrey, D., Perlman, J., Marshall, G.N. and Hambarsoomians, K. (2009). Use of expert ratings as sampling strata for a more cost-effective probability sample of a rare population. *Public Opinion Quarterly*, 73, 56-73.
- Erens, B., Prior, G., Korovessis, C., Calderwood, L., Brookes, M. and Primatesta, P. (2001). Survey methodology and response. In *Health Survey for England – The Health of Minority Ethnic Groups '99. Volume 2: Methodology and Documentation*. (Eds., B. Erens, P. Primatesta and G. Prior). The Stationery Office, London.
- Fecso, R.S., Baskin, R., Chu, A., Gray, C., Kalton, G. and Phelps, R. (2007). *Design Options for SESTAT for the Current Decade*. Working Paper SRS 07-021. Division of Science Resource Statistics, U.S. National Science Foundation.
- Fiscella, K., and Fremont, A.M. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health Services Research*, 41, 1482-1500.
- Flores Cervantes, I., and Kalton, G. (2008). Methods for sampling rare populations in telephone surveys. In *Advances in Telephone Survey Methodology*. (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japac, P. Lavrakas, M.W. Link and R.L. Sangster). Hoboken, NJ: Wiley, 113-132.
- Folsom, R.E., Potter, F.J. and Williams, S.K. (1987). Notes on a composite size measure for self-weighting samples in multiple domains. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 792-796.
- German Federal Statistical Office (2009). *Microcensus*. Available at http://www.destatis.de/jetspeed/portal/cms/Sites/destatis/Internet/EN/press/abisz/Mikrozensus__e.templateId=renderPrint.psmml.
- Goldman, J.D., Borrud, L.G. and Berlin, M. (1997). An overview of the USDA's 1994-96 Continuing Survey of Food Intakes by Individuals and the Diet and Health Knowledge Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 796-801.
- Gonzalez, J.F., Ezzati, T.M., White, A.A., Massey, J.T., Lago, J. and Waksberg, J. (1985). Sample design and estimation procedures. In *Plan and Operation of the Hispanic Health and Nutrition Examination Survey, 1982-84*. (Ed., K.R. Maurer). Vital and Health Statistics, Series 1, No. 19. U.S. Government Printing Office, Washington, DC, 23-32.
- Green, J. (2000). Mathematical programming for sample design and allocation problems. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 688-692.
- Haerer, A.F., Anderson, D.W. and Schoenberg, B.S. (1986). Prevalence and clinical features of epilepsy in a biracial United States population. *Epilepsia*, 27, 66-75.
- Hartley, H.O. (1974). Multiple frame methodology and selected applications. *Sankhyā*, 36, 99-118.
- Heckathorn, D.D. (1997). Respondent driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44, 174-199.
- Heckathorn, D.D. (2007). Extensions of respondent-driven sampling: Analyzing continuous variables and controlling for differential recruitment. *Sociological Methodology*, 37, 151-208.
- Hedges, B.M. (1973). *Sampling Minority Groups*. Thomson Medal Awards. Thomson Organization, London.
- Hedges, B.M., (1979). Sampling minority populations. In *Social and Educational Research in Action* (Ed., M.J. Wilson) London: Longman, 245-261.
- Horrigan, M., Moore, W., Pedlow, S. and Wolter, K. (1999). Undercoverage in a large national screening survey for youths. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 570-575.
- Kalsbeek, W.D. (2003). Sampling minority groups in health surveys. *Statistics in Medicine*, 22, 1527-1549.
- Kalton, G. (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17, 183-194.
- Kalton, G. (1993a). *Sampling Rare and Elusive Populations*. Department for Economic and Social Information and Policy Analysis, United Nations, New York.

- Kalton, G. (1993b). Sampling considerations in research on HIV risk and illness. In *Methodological Issues in AIDS Behavioral Research*. (Eds., D.G. Ostrow, R.C. Kessler). New York: Plenum Press.
- Kalton, G. (2003). Practical methods for sampling rare and mobile populations. *Statistics in Transition*, 6, 491-501.
- Kalton, G., and Anderson, D.W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society, A*, 149, 65-82.
- Kalton, G., and Brick, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.
- Kanouse, D.E., Berry, S.H. and Duan, N. (1999). Drawing a probability sample of female street prostitutes in Los Angeles County. *Journal of Sex Research*, 36, 45-51.
- Katzoff, M.J. (2004). Applications of adaptive sampling procedures to problems in public health. In *Proceedings of Statistics Canada Symposium 2004, Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at <http://www.statcan.gc.ca/pub/11-522-x/2004001/8751-eng.pdf>
- Katzoff, M.J., Sirken, M.G. and Thompson, S.K. (2002). Proposals for adaptive and link-tracing sampling designs in health surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1772-1775.
- Kish, L. (1965a). Selection techniques for rare traits. In *Genetics and the Epidemiology of Chronic Diseases*. Public Health Service Publication No. 1163.
- Kish, L. (1965b). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Kish, L. (1976). Optima and proxima in linear sample design. *Journal of the Royal Statistical Society, A*, 139, 80-95.
- Kish, L. (1987). *Statistical Design for Research*. New York: John Wiley & Sons, Inc.
- Kish, L. (1988). Multipurpose sample design. *Survey Methodology*, 14, 19-32.
- Kish, L. (1999). Cumulating/combining population surveys. *Survey Methodology*, 25, 129-138.
- Körner, T., and Nimmergut, A. (2004). A permanent sample as a sampling frame for difficult-to-reach populations? In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at <http://www.statcan.gc.ca/pub/11-522-x/2004001/8752-eng.pdf>.
- Langa, K.M., Plassman, B.L., Wallace, R.B., Herzog, A.R., Heeringa, S.G., Ofstedal, M.B., Burke, J.R., Fisher, G.G., Fultz, N.H., Hurd, M.D., Potter, G.G., Rodgers, W.L., Steffens, D.C., Weir, D.R. and Willis, R.J. (2005). The Aging, Demographics, and Memory Study: Study design and methods. *Neuroepidemiology*, 25, 181-191.
- Lauderdale, D.S., and Kestenbaum, B. (2000). Asian American ethnic identification by surname. *Population Research and Policy Review*, 19, 283-300.
- Lavallée, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- Lavallée, P. (2007). *Indirect Sampling*. New York: Springer.
- Lohr, S.L. (2009). Multiple-frame surveys. In *Handbook of Statistics. Volume 29A: Sample Surveys: Design, Methods, and Applications*. (Eds., D. Pfeffermann and C.R. Rao). Burlington, MA: Elsevier B.V.
- Lohr, S.L., and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.
- Lohr, S.L., and Rao, J.N.K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101, 1019-1030.
- Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 87-96.
- MacKellar, D., Valleroy, L., Karon, J., Lemp, G. and Janssen, R. (1996). The Young Men's Survey: Methods for estimating HIV seroprevalence and risk factors among young men who have sex with men. *Public Health Reports*, 111, Supplement 1, 138-144.
- Maffeo, C., Frey, W. and Kalton, G. (2000). Survey design and data collection in the Disability Evaluation Study. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 79-88.
- Marker, D.A. (2001). Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators. *Survey Methodology*, 27, 183-188.
- McKenzie, D.J., and Mistiaen, J. (2009). Surveying migrant households: A comparison of census-based, snowball and intercept point surveys. *Journal of the Royal Statistical Society*, 172, 339-360.
- Meccati, F. (2004). Center sampling: A strategy for sampling difficult-to-sample populations. In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at <http://www.statcan.gc.ca/pub/11-522-x/2004001/8740-eng.pdf>.
- Metcalfe, P., and Scott, A. (2009). Using multiple frames in health surveys. *Statistics in Medicine*, 28, 1512-1523.
- Michael, R.T., and O'Muircheartaigh, C.A. (2008). Design priorities and disciplinary perspectives: The case of the US National Children's Study. *Journal of the Royal Statistical Society, A*, 171, 465-480.
- Mohadjer, L., and Curtin, L.R. (2008). Balancing sample design goals for the National Health and Nutrition Examination Survey. *Survey Methodology*, 34, 119-126.
- Morris, P. (1965). *Prisoners and Their Families*. Allen and Unwin, London, 303-306.
- National Children's Study (2007). Study design. In *The National Children's Study Research Plan*. Version 1.3. Available at http://www.nationalchildrensstudy.gov/research/studydesign/researchplan/Pages/Chapter_6_032008.pdf
- Pew Research Center (2007). *Muslim Americans, Middle Class and Mostly Mainstream*. Pew Research Center, Washington, DC.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Rodriguez Vera, A. (1982). Multipurpose optimal sample allocation using mathematical programming. Biostatistics Doctoral Dissertation. Ann Arbor: University of Michigan.

- Rothbart, G.S., Fine, M. and Sudman, S. (1982). On finding and interviewing the needles in the haystack: The use of multiplicity sampling. *Public Opinion Quarterly*, 46, 408-421.
- Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 3-22.
- Sirken, M.G. (2004). Network sample surveys of rare and elusive populations: A historical review. In *Proceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to-Reach Populations*. Available at <http://www.statcan.gc.ca/pub/11-522-x/2004001/8614-eng.pdf>.
- Sirken, M.G. (2005). Network sampling developments in survey research during the past 40+ years. *Survey Research*, 36, 1, 1-5. Available at <http://www.srl.uic.edu/Publist/Newsletter/pastissues.htm>.
- Skinner, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.
- Smith, P.J., Battaglia, M.P., Huggins, V.J., Hoaglin, D.C., Roden, A., Khare, M., Ezzati-Rice, T.M. and Wright, R.A. (2001). Overview of the sampling design and statistical methods used in the National Immunization Survey. *American Journal of Preventive Medicine*, 20(4S), 17-24.
- Statistics Canada (2008). *Canadian Community Health Survey (CCHS)*. Available at <http://www.statcan.gc.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3226&lang=en&db=imdb&adm=8&dis=2#b3>.
- Sudman, S. (1972). On sampling of very rare human populations. *Journal of the American Statistical Association*, 67, 335-339.
- Sudman, S. (1976). *Applied Sampling*. New York: Academic Press.
- Sudman, S., and Freeman, H.E. (1988). The use of network sampling for locating the seriously ill. *Medical Care*, 26, 992-999.
- Sudman, S., and Kalton, G. (1986). New developments in the sampling of special populations. *Annual Review of Sociology*, 12, 401-429.
- Sudman, S., Sirken, M.G. and Cowan, C.D. (1988). Sampling rare and elusive populations. *Science*, 240, 991-996.
- Thompson, S.K. (2002). *Sampling*. 2nd Edition. New York: John Wiley & Sons, Inc.
- Thompson, S.K., and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, 26, 87-98.
- Thompson, S.K., and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: John Wiley & Sons, Inc.
- Tortora, R., Groves, R.M. and Peytcheva, E. (2008). Multiplicity-based sampling for the mobile telephone population: Coverage, nonresponse, and measurement issues. In *Advances in Telephone Survey Methodology*. (Eds., J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japec, P.J. Lavrakas, M.W. Link and R.L. Sangster). Hoboken, NJ: Wiley, 133-148.
- U.S. Census Bureau (2009a). *Design and Methodology, American Community Survey*. U.S. Government Printing Office, Washington, DC.
- U.S. Census Bureau (2009b). *Small Area Income and Poverty Estimates*. Available at <http://www.census.gov/did/www/saiper/methods/statecounty/index.html>.
- U.S. National Center for Health Statistics (2009a). *National Health Interview Survey (NHIS)*. Available at <http://www.cdc.gov/nchs/nhis/methods.htm>.
- U.S. National Center for Health Statistics (2009b). *The National Immunization Survey (NIS)*. Available at http://www.cdc.gov/nis/about_eng.htm.
- U.S. National Center for Health Statistics (2009c). *State and Local Area Integrated Telephone Survey (SLAITS)*. Available at <http://www.cdc.gov/nchs/about/major/slaits/nsch.htm>.
- Volz, E., and Heckathorn, D.D. (2008). Probability based estimation theory for respondent driven sampling. *Journal of Official Statistics*, 24, 79-97.
- Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. *Proceedings of the Social Statistics Section*, American Statistical Association, 429-434.
- Waksberg, J., Brick, J.M., Shapiro, G., Flores Cervantes, I. and Bell, B. (1997). Dual-frame RDD and area sample for household survey with particular focus on low-income population. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 713-718.
- Waksberg, J., Judkins, D. and Massey, J.T. (1997). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology*, 23, 61-71.
- Watters, J.K., and Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, 36, 416-430.
- Winglee, M., Park, I., Rust, K., Liu, B. and Shapiro, G. (2007). A case study in dual-frame estimation methods. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3195-3202.
- Word, D.L., and Perkins, R.C. (1996). *Building a Spanish Surname List for the 1990's - A New Approach to an Old Problem*. Population Division Technical Working Paper No. 13. U.S. Census Bureau, Washington, DC.
- Wu, C., and Rao, J.N.K. (2009). Empirical likelihood methods for inference from multiple frame surveys. *Proceedings of the International Statistical Institute*, Durban, South Africa.
- Xia, Q., Tholandi, M., Osmond, D.H., Pollack, L.M., Zhou, W., Ruiz, J.D. and Catania, J.A. (2006). The effect of venue sampling on estimates of HIV prevalence and sexual risk behaviors in men who have sex with men. *Sexually Transmitted Diseases*, 33, 545-550.