INITIAL SAMPLE DESIGN CONSIDERATIONS FOR THE NATIONAL CHILDREN'S STUDY[1]

David Judkins, Graham Kalton, Joseph Waksberg, Westat; John Kiely, Amy Branum, NCHS; and Peter Scheidt, NICHD
Westat, 1650 Research Boulevard, Rockville, Maryland 20850

Key Words: Antenatal Sampling, Gravida Recruitment, Pregnancy Screening

1. Introduction

The Children's Health Act of 2000 (PL 106-310) mandated the National Institute of Child Health and Human Development (NICHD), the Environmental Protection Agency (EPA), and the Centers for Disease Control and Prevention (CDC), to plan, develop, and conduct a prospective cohort study, now known as the National Children's Study (NCS). The goals of this study are to address important medical, environmental, and social hypotheses concerning the effects of environmental exposures on children's physiology, emotional development, and cognitive abilities. The NCS will be national in scope and will involve the study of approximately 100,000 children from a point in time before birth through age 21. The proposed sample size of 100,000 may be revised later as detailed plans for the NCS evolve. The NCS data collections will include biologic data, environmental exposure data, and neighborhood data, as well as personal information reported by mothers and children. One goal is that the study's findings should be generalizable as closely as possible to the population of the United States as a whole. The study should also include sufficient sample sizes of children in a sizable number of groups of particular interest to produce separate reliable estimates for these groups.

Under contract to the National Center for Health Statistics (NCHS), Westat was tasked with developing and evaluating a number of candidate sample frames and sample designs for NCS enrollment. Members of NICHD and NCHS participated in this work through guidance, discussion and reviews. In this paper we mostly summarize a more detailed Westat report (Westat, 2002).

The team developed three sampling models for initial consideration: the Household Model (door-to-door screening for fecund women), the Office Model (recruitment of pregnant women during ordinary prenatal care visits), and the Center Model (recruitment of pregnant women through a small number of formal centers that would be responsible for executing all aspects of the study protocol for their own recruits throughout the life of the project). Two variants of the Household Model with different degrees of clustering were examined. Thus, we evaluated these four designs. For this preliminary design work, no allowance was made for any oversampling of subgroups of children of special interest.

The full report discusses the type and degree of clustering in the four evaluated designs, initial sample size determination, detailed costs for the sample recruitment, some aspects of the relative difficulty of various measurements of exposure and outcomes under the alternative designs, and statistical power for various tests. From this initial study of sample design options, this paper describes general considerations in frame choice and level of clustering (Section 2), operations involved in each sampling procedure (Sections 3 through 5), a summary of the advantages and disadvantages of the different designs (Section 6), a brief note on hybrid sampling models (Section 7), and some caveats (Section 8).

2. Frame and Clustering Considerations

Frame Selection

As a rule, large-scale national surveys employ probability sample designs so that the survey results will reflect data for the total population using statistical methods that do not depend on untestable assumptions. However, it must be born in mind that, as a long-term longitudinal cohort study with complex exposure and outcome measures, sampling considerations are more difficult and problematic than for a typical survey. Probability sample designs involve the identification of a well-defined sampling frame and then of methods to select a sample from that frame. The Household and Office Models employ probability sampling designs. Since there are concerns about the feasibility and costs involved in selecting a national probability sample of

[1] This research was carried out under contract HHS-100-97-0017 with the National Center for Health Statistics, which was funded by Interagency Agreement HD-128401 from NICHD and the Interagency Coordinating Committee of the National Children's Study.

pregnant women for the NCS, we also considered a nonprobability approach and associated model for this purpose. The Center Model is a nonprobability design.

An important assumption underlying the choice of these models is a requirement to begin health and related measurements during pregnancy. Most of the cost of sample selection would be eliminated and a more efficient sample could be selected if the first measurements could be delayed until approximately 6 to 12 months after birth. This could be accomplished by selecting, whenever possible, a sample of the births reported in the Birth Registration System. (This sampling method is used in the National Center for Education Statistic's Early Childhood Longitudinal Study—Birth Cohort, commonly referred to as ECLS-B.) A sample design based on this frame would provide better coverage, and probably better response rates, than the designs studied here. Also, a sample design based on this frame can readily and economically oversample any subgroups of analytic interest for which subgroup membership is recorded on the birth certificate records. However, the NCS requirement that broad environmental exposures, health measurements, and certain other data be collected as early in pregnancy as possible rule out a birth registration sample.

Another sample design that had to be ruled out was one in which women are sampled in delivery rooms. This design would involve a probability sample of hospitals, birthing centers, etc., and a probability sample of women giving birth in the sampled locations. This type of design would be an attractive one for the collection of placentas and cord blood samples. It was, however, excluded from consideration because it fails to provide the required prenatal data.

Clustering

Each of these models assumes that a clustered sample of pregnant women is selected. The use of such clustering substantially reduces survey costs and aids operational feasibility. However, it lowers the precision of survey estimates for a given sample size: the greater the clustering of the sample, the greater the loss of precision. The ideal practice is to take both cost and precision into account to produce an "optimum sample design," i.e., one in which the amount of clustering produces the lowest sampling error for a given estimate for a fixed total cost. This ideal approach, however, cannot be applied exactly to multi-purpose surveys unless one specific statistic (such as the unemployment rate) is given priority over all other statistics.

No single sample design can provide optimal clustering for all statistics for a survey with as many objectives as the NCS. The efficiency of a particular clustering plan will depend on a variety of factors, mainly the items of analysis, the importance of information on risk factors that affect all children vs. the effects on subgroups (such as minorities, children in rural areas, those living in areas subject to particular environmental conditions, those growing up in households with various social or economic backgrounds, etc.), the effects of clustering on the cost of data collection, and the degree to which clustering affects the ability to carry out followup activities for persons who move from their initial sample location.

An important feature of any analysis of NCS data is that it should take the clustering of the sample into account. With complex probability sample designs, such as those used in the Household and Office Models, this can be done in a standard way using one of the various software packages for survey analysis that take account of the sample design, and in particular the clustering involved. The precision of estimates and the power of significance tests with data from nonprobability designs like that used in the Center Model are also affected by clustering, and that needs to be reflected in the analytic methods employed. Often the approach used in this kind of case is to employ some form of random effects model—hierarchical or multilevel model—to take account of the clusters (Bryk and Raudenbush, 1992; Longford, 1993; Goldstein, 1995; Hox, 2002). The application of such models is now straightforward given the availability of several packages for computing them. It is important to use such methods in order to avoid the overstatement of the precision of estimates and the increased false positive results in significance tests that result from the use of standard methods that fail to reflect the clustering. It is of interest to note here that the same issue of clustering arises in multi-center clinical trials; the need to take the clustering into account in the analysis was noted many years ago by Cornfield (1978) and is emphasized in the recent book by Donner and Klar (2000). The effect of taking the clustering into account through multilevel models is similar to that obtained by applying the survey sampling approach. It results in a design effect like that for clustering developed in the survey sampling literature (e.g., Kish, 1965). We have therefore applied a design effect with the Center Model in the same way as with the other models.

For simple estimation of the marginal distributions of exposure and health outcomes, the effect of two levels of clustering can be expressed in terms of the design effect (DE) given by

$$\mathrm{DE} = 1 + \delta_1(\bar{n} - 1) + \delta_2(\bar{\bar{n}} - 1)$$

where δ_1 and δ_2 are the intraclass correlation coefficients that measure the homogeneity of the condition of interest at the two levels, and where $\bar{n}$ and

$\bar{\bar{n}}$ are the average sample sizes per first- and second-stage cluster respectively.

Given the immense size of the NCS with 100,000 children, it is clear that even modest intraclass correlation can lead to very large design effects with the resulting loss of precision if the survey is confined to a small number of clusters. For example, we estimated that the design effect for a Center Model with 60 centers would be on the order of 18 for estimates relating to the full sample, implying that the effective sample size would be only 5,600. Of course, estimation of marginal exposure and outcome distributions are not the focus of the NCS. Rather, with it's 20 years of followup, the central focus of NCS is on the relationships of early exposures to later outcomes. The impact of clustering on the precision of analytic statistics such as regression coefficients and relative risks is lower than that on marginal estimates, but nevertheless can still be substantial with a highly clustered sample. Design effects on analytic statistics larger than one can arise because of unmeasured factors that affect the relationship and that vary across clusters. (See for example, Kalton and Blunden, 1973; Kish and Frankel, 1974; Holt and Scott, 1981; Scott and Holt, 1982; and Wu, Holt and Holmes, 1988.)

Thus, from the standpoint of precision, there is an important advantage for spreading the NCS across a large number of clusters, perhaps on the order of 800. However, such a highly dispersed sample is a disadvantage for the NCS in terms of the complexity of the health outcome measures that can be attempted. Those measures that would require expertise of staff at academic medical centers would be difficult to perform on a widely dispersed sample, requiring extensive travel on the part of respondents or the transportation of mobile medical centers close to subject homes as is done in the National Health and Nutrition Examination Survey (NHANES). However, the measurement of many health outcomes does not require such expertise. In the full report, we outline possible strategies for using survey interviewers, home nurses, a network of preferred providers, or a network of survey-established medical offices to conduct the measurements.

A final note on the implications of clustering is that the originally selected sample will become less clustered because of migration. The impact of migration on data collection will obviously be more striking in a design that is initially tightly clustered than in one that is initially dispersed. Protocols that are initially designed for a dispersed sample will stand up better over time.

The choice of model somewhat constrains the range of clustering that can be considered. For both the Household and Office Models, we envisioned counties as the primary sampling units (PSUs) for the first stage of selection. We considered a range between 100 and 800 sample counties. The second stage of selection would be "segments" of neighboring blocks for the Household Model and doctors' offices for the Office Model. We considered a range of 1,000 to 50,000 segments for the Household Model and 500 to 4,000 offices for the Office Model. For the Center Model, the problems of design effects, on the one hand, and coordination and adequate work per center on the other hand, led us to consider a range of 60 to 100 centers. The designs we studied in detail are summarized in Table 2 in Section 6.

3. The Household Model

The basic idea of the Household Model is to use door-to-door screening of a nationally representative sample households to find fecund women and then to recruit these women into a screening panel which is periodically recontacted to discover new pregnancies. Once pregnancies are reported, the women would be recruited into the NCS. Table 1 shows the sample sizes at each phase.

Table 1. Preliminary estimates of sample sizes for Household Model*

Phase	Sample count
Initial sample of listed dwelling units	1,317,000
Households (residential occupied dwelling units)	1,159,000
Screened households	1,101,000
Age-eligible women and girls	622,000
Age-eligible females screened forcurrent pregnancy	498,000
Responding women not surgically sterile	364,000
Number current pregnancies initially reported	16,000
Number additional pregnancies reported over the following 35 months	102,000
Total pregnancies reported and baseline interviews given	118,000
Infant exams on live births	100,000

*More work is required to fine-tune these numbers. Research is needed on the coverage that is likely to be achieved of intended and unintended pregnancies and abortion rates.

The initial sample of dwelling units would be selected by standard area-sampling techniques as

described in Hansen, Hurwitz, and Madow (1953). Trained interviewers would visit each household to determine whether there are any female occupants in their child-bearing years (restricted to ages 15 through 44, following the tradition in the National Survey of Family Growth) and female occupants who will enter child-bearing years in the next 36 months (thus reducing the lower age limit to 12). Interviewers could collect the data from any responsible member of the household. They would also be allowed to ask neighbors as a last resort. Based on similar experience in other surveys, we think it is possible to achieve a response rate of 95 percent for this phase of the operation.

We considered two variants with different degrees of clustering. One would have 12,500 segments (on the order of census block groups) in 800 counties. The other would have 3,125 segments (on the order of census tracts) in 300 counties. One reason to consider the smaller number of segments is that it would facilitate the measurement of neighborhood-level environmental exposures.

Once the eligible households had been identified, female interviewers would visit each household and administer a brief screening questionnaire (about 10 minutes) to all female occupants aged 15-44 years old on surgical sterilization and current pregnancy. We believe that is important at this phase to conduct personal interviews with the women in a private setting, shielding their replies from other household members. We think that with a $10 incentive, it would be possible to achieve a response rate of 80 percent for these screening interviews. Based on pregnancy data from the National Center for Health Statistics and certain assumptions about the forthrightness of respondents, we estimate that 16,000 current pregnancies would be reported. To obtain the balance of the sample, the women who were not surgically sterile would be recontacted every three months. These contacts would mostly be made by phone, but personal visits would be made to those women without phone service in order to keep the sample as representative as possible.

As pregnancies are reported, the women would be asked to participate in 45-minute baseline interviews with an incentive of $50 per interview. Of course, some of the initial respondents will drop out of the study before reporting pregnancies, others will conceal their pregnancies from the interviewers, and even among those who reveal their pregnancies, not all will consent to a baseline interview. Taking these factors into account, we estimate that three years would be adequate to find and conduct baseline interviews with 118,000 pregnant women. If this number of interviews were not achieved after three years, it would be simple to extend the length of the screening period.

Given the focus of the study, we assume that most women planning to seek abortions will not report their pregnancies. Also, given that most miscarriages occur early in pregnancy, we assumed that 90 percent of the reported pregnancies with baseline interviews would result in live births. Based on infant examination rates in NHANES, we estimated that 94 percent of the newborns would be examined. These assumptions lead to the projection of about 100,000 infant exams.

4. The Office Model

The basic idea of the Office Model is to utilize the existing infrastructure of prenatal care providers to quickly and inexpensively recruit a highly dispersed national probability sample of pregnant women. The design we studied involves recruiting 4,000 offices in 800 counties across the nation. As in the Household Model, there would be several phases in the sampling procedure. The first phase would be to conduct a prescreening survey of 40,800 prenatal care providers across the 800 sample counties using lists of such providers obtained from the American Medical Association. (In many counties, this may imply taking a census of these providers.) Based on research by Kalsbeek and Mancewicz (1993), we would recommend restricting the census to physicians with general practice, family practice, obstetrics-gynecology, obstetrics, gynecology, maternal-fetal medicine, reproductive endocrinology, or infertility as a primary or secondary specialty. The point of this census would be to identify all ancillary practice locations of the providers and to further estimate the volume of prenatal care services delivered by each provider at each practice location. Physicians would be asked to report services provided by midwives under their direction as separate locations. We project a 70 percent response rate on the prescreening.

At the second phase, a subsample of 5,700 provider/location combinations (“offices” in the balance of the paper) would then be selected for recruitment. The duties of recruited offices would include patient recruitment as well as representing the study before any local Institutional Review Board (IRB). We project a 70 percent recruitment rate, resulting in a sample of 4,000 cooperating offices. This projection assumes that payments would be made to the offices as well as to staff working at them. We allowed for a combined payment of about $6,400 per cooperating office, with an additional allowance of $1,000 for the purchase of computer hardware and/or software.

At the third phase, sample providers would be asked to recruit all their new prenatal care patients at the sample practice location for a fixed number of weeks, depending on the previously estimated volume of the provider at that location. The recruitment window

would be short at high volume offices and long at low-volume offices.

During the target weeks, all women seeking their first prenatal care would be in sample. For simplicity and for better control on the process, there would be no subsampling within the weeks. In those obstetrical offices where the patients are deliberately rotated among the physicians in the practice, we would associate women on the first prenatal care visit with the physician that they happened to see on that visit.

Most of the recruiting work would be done by doctors and their staffs, although it might be practical to use professional recruiters in large offices. The recruiter would be required to keep track of both recruitment successes and failures. Additional work would be needed to develop and test the best recruitment procedures, but one reasonable model might be to have the doctor briefly mention to women that either one of his nurses or a professional recruiter will see them next to discuss participation in an important study. It might also be sensible to have the recruiting physicians carry out whatever physical examinations are required of the pregnant women.[2]

Although doctors collect their own case histories, it will be important to have pregnant women report their histories as part of the survey's standardized baseline questionnaire. One approach for administering the baseline questionnaire would be to have the women complete the questionnaires on their first visits using a self-completion paper-and-pencil or computer-based questionnaire. If a computer-based system were used, the NCS could support the physicians and their office staff by supplying them with training manuals, computers, software, information brochures, and a help line. However, if the initial round of data collection is a detailed one, the work to train the staff in 4,000 offices to manage the data collection in an effective way seemed too daunting.

An alternative approach would be to have the sampled women give consent to having their names and addresses being provided to a survey organization, which would then send interviewers to their homes to conduct the baseline interviews. Telephone appointments could be made to make the personal visits more efficient. The self-completion of the baseline questionnaire in the office is an untried method and its success is uncertain. We have therefore assumed for the Office Model that the alternative approach would be used of conducting baseline interviews by survey interviewers in the sampled women's homes. However, the need for a subsequent visit to obtain the baseline information will result in some additional refusals. Women completing the baseline interview would be given a $50 incentive, as in the Household Model.

A concern with the office-based enrollment is the difficulty of controlling the selection and recruitment of the pregnant women. These operations will be conducted without the tight supervisory control that is applied in standard survey settings. Moreover, they would often be left to office staff who lack the training and commitment of professional survey interviewers. Consequently, there is a serious risk that some women who should be sampled will be missed and many others will not be persuaded to participate in the NCS.

We estimate a response rate of about 70 percent in the prescreening of physicians and a 70 percent recruitment rate among those subsampled after prescreening. The recruitment rate for pregnant women that might be achieved within sampled offices is very uncertain, but a recent study in Denmark that used procedures similar to those proposed here for the office model obtained a recruitment rate of 60 percent of patients at participating doctors (Olsen, et al, 2001). Multiply these three rates together gives an overall response rate of 29 percent. We note that the Danish study obtained a doctor recruitment of 60 percent, somewhat higher than the 50 percent we are estimating.

5. The Center Model

The Center Model involves the selection and funding of a small number of large health care centers, each of which would be responsible for recruiting its sample of pregnant women. To be manageable, the number of centers would need to be relatively small, but the design effects are severe for even as many as 100 centers. We studied a design with 100 centers in the belief that this is close to the maximum feasible. Based on 100 centers, on average each center would be expected to recruit about 1,200 pregnant women over a three to five year period. The expectation is that the centers would then be responsible for the ongoing measurements and health examinations and for maintaining current records for the children resulting from live births for the duration of the NCS. A coordinating center would be needed to ensure common procedures and comparability of data between the centers. Each center would no doubt have its own IRB, and the coordinating center would need to assure that the standardized procedures satisfy the requirements of all the individual IRBs. Questions of data rights for the center directors might also be a complicating factor.

[2] Note, however, that the costs presented in Table 2 do not include the costs for implementing this suggestion.

Some sort of contracting or grant awarding process would be required to establish the network of centers. Bidders might include teaching hospitals, other hospitals, HMOs, rural health clinics, and other clinics. This could be done by the coordinating center or directly by the government. Centers would need to have a charter, space, personnel, and a model for recruiting pregnant women. Presumably, there would be goals set for the geographic dispersion of the centers. There might also be goals set for the mix of patients with respect to race and socio-economic class.

6. Comparing the Designs

At the time of this initial sample design study, hypotheses were being developed to serve both as justification for the NCS and as guidance for which design to chose. This work of developing the hypotheses continues. Some of the hypotheses emphasize the importance of measuring various exposure levels near the time of conception. These hypotheses are best supported by the Household Model since this model provides for the earliest recruitment and can be easily modified to recruit women prior to the conception if necessary (but at considerable cost). Other hypotheses emphasize placental and neonatal infections, for which the collection of placentas and/or cord blood samples would be very important—perhaps even essential. Although this issue was raised too late in our study to receive careful attention, it appears that these hypotheses would be best supported by the Center Model, where the personal investment of obstetricians and their institutions would facilitate the collection of biologic specimens. Yet other hypotheses emphasize the relationship between early childhood social environment and various mental health outcomes. These hypotheses would probably be best served by a birth-certificate sample or a sample of births. Thus, it is difficult to make firm recommendations about the best sample design until the relative importance of the various hypotheses has been settled.

Nonetheless, some comparisons can be made. These are summarized in Table 2. The reader is referred to the full report for the reasoning behind these assessments. A few points deserve closer attention.

- With respect to the initial costs (project set-up operations, sampling, recruitment, and baseline maternal interviews), the difference between the least and most expensive options is less than three percent of the anticipated $3 billion cost of the NCS.
- Oversampling targeted groups was outside of the scope of the design study, but it is clear that oversampling of targeted groups is likely to be very expensive with any design. Oversampling on the basis of perceived risk is particularly problematic since it assumes strong hypotheses about risks and since many putative risk factors are likely to be more a question of personal lifestyles than of geography. Also, oversampling of some subgroups is at the expense of a reduction in sample size for other subgroups. The precision of estimates that do not depend on membership of the oversampled subgroup will be lower.
- Response rates are not as high as can be achieved on simpler surveys such as those on social-economic conditions. Nonetheless, there is a major difference between a probability design with high nonresponse and a nonprobability design; in the former, something is known about the people who declined to participate whereas in the latter, the inevitable biases in the selections are unknown.

7. Hybrids

Reviewers of the initial draft of this full report suggested consideration of various "hybrid" designs. The main point of these hybrids[3] was to try to combine the benefits of the Household Model, such as generalizability and recruitment early in pregnancy, with the benefits of the Center Model such as administrative efficiency and the collection of biologic specimens (e.g., placentas) that must be made in hospitals. A significant limitation to this approach is that the number of PSUs probably has to be limited to be not much greater than 100 since each participating center would probably only be able to supervise local operations. Moreover, appropriate centers would then need to be found for the sample PSUs. As we have noted elsewhere, there are significant analytic advantages attached to the use of a large number of PSUs to locate pregnant women. A hybrid design would require giving up those analytic advantages. The ultimate decision on the number of PSUs will depend on balancing these two considerations.

[3] One of these was essentially the Household Model with 50 PSUs instead of 300 or 800. Another was to use the Center Model but encourage the centers to employ household screening. This, too, is very similar to the Household Model, but with a decentralized organizational structure.

Table 2. Design comparisons

Feature	Household Model A	Household Model B	Office Model	Center Model
Type of first-stage cluster	County	County	County	Center
Number of first-stage clusters	800	300	800	100
Type of second-stage cluster	Groups of about 5 neighboring census blocks	Groups of about 21 neighboring blocks	Office (provide/location)	NA
Number of first-stage clusters	12,500	3,125	4,000	NA
Probability Design?	Yes	Yes	Yes	No
Coverage of mothers who never seek prenatal care?	Yes	Yes	No	No
Cumulative response rate through baseline maternal interview	68%	68%	29%	NA
Ranking for lowest gestational age at enrollment (1=best, 4=worst)	1	1	3	4
Could be adapted to enroll women prior to conception?	Yes	Yes	No	Doubtful
Ranking for ease of collection of birth biologics (1=best, 4=worst)	4	3	2	1
Ranking for ease of collection of home environmental samples such as dust and water (1=best, 4=worst)	1	1	1	4
Ranking for ease of collection of neighborhood social capital other than census data (1=best, 4=worst)	2	1	3	4
Likely design effects for marginal estimates	2.5	11.5	3.0	11.0
Power to detect association when exposure prevalence is 20 percent, disease prevalence is 0.2 percent and relative risk is 2.0	71-83%	49-70%	65-81%	32-50%
Estimated cost to set up system, draw sample, and conduct baseline interviews with pregnant women (millions)	189	163	109	166
Ranking for speed of recruiting (1=fastest, 4=slowest)	3	3	1	2
Ranking for ability to enforce rules for protection of human subjects during recruiting (1=best, 4=worst)	2	2	4	1
Ranking for ability to minimize measurement variance on health outcomes (1=best, 4=worst)	3	2	3	1
Requirement to work through local gatekeepers?	No	No	Yes	Yes
Precedents?	No	No	No	Yes
Ability to adapt to oversample racial and ethnic domains	Fair	Fair	Poor	Fair
Ability to adapt to oversample rare groups with high exposure levels that are believed to place their children at high risk	Poor	Poor	Poor	Poor
Robustness to migration	Excellent	Good	Excellent	Fair

A secondary point of such hybrid designs is that they likely lead to decentralized control of the study. A decentralized structure might be able to better assemble and utilize a variety of resources as well as being more responsive to local issues. However, centralized hierarchical structure for all operations should result in the most uniform measurements. Each of the major types of data collection could be assigned to different agents in a centralized hierarchical structure. It would also be possible to use a centralized hierarchical structure for the recruitment phase while using a decentralized structure for all the other measurements provided that the number of PSUs was small enough, as indicated above.

8. Caveats

The observations and statements made in this report are based on experience and judgments of the authors and a number of assumptions of the Westat staff and the staff of the Federal agencies engaged in planning the study. As such they are subject to change and should not be taken as given or fact. This examination of sampling options highlighted the importance of reaching consensus on the most important hypotheses for the NCS.

9. References

Bryk, A.S. and Raudenbush, S.W. (1992). *Hiearchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Cornfield, J. (1978). Randomization by group, a formal analysis. *American Journal of Epidemiology*, 108, 100-102.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research.* London: Arnold.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd ed. London: Edward Arnold.

Hansen, M.H., Hurwitz, W.N., and Madow, W. G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley.

Holt, D. and Scott, A.J. (1981). Regression analysis using survey data. *The Statistician,* 30, 169-178.

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. Mawah, NJ: Lawrence Erlbaum Associates.

Kalsbeek, W.D. and Macewicz, M.J. (1993). Sampling prenatal care providers from a frame of physicians. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 206-211. Alexandria, VA: American Statistical Association.

Kalton, G. and Blunden, R.M. (1973). Sampling errors in the British General Household Survey. *Bulletin of the International Statistical Institute, Book 3 of the 1973 Proceedings*, 83-97.

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

Kish, L. and Frankel, M.F. (1974). Inference from Complex Samples (with discussion). *Journal of the Royal Statistical Society, Series B,* 36, 1-37.

Longford, N.T. (1993). *Random Coefficient Models.* Oxford: Clarendon Press.

Scott, A.J. and Holt, D. (1982). The effects of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association,* 77, 848-854.

Olsen, J., Melbye, M., Olsen, S.F., Sorensen, T. I., Aaby, P., Andersen, A.M., Taxbol, D., Hansen, K. D., Juhl, M., Schow, T.B., Sorensen, H.T., Andresen, J., Mortensen, E L., Olesen, A.W. and Sondergaard, C. (2001). The Danish National Birth Cohort—its background, structure and aim. *Scand J Public Health*. 29, 300-307.

Westat (2002). Sampling strategies for the proposed National Children's Study. Final report. (Prepared under contract to the Department of Health and Human Services). Rockville, MD: Westat, Inc.

Wu, C.F.J., Holt, D. and Holmes, D.J. (1988). The effect of two-stage sampling on the F statistic. *Journal of the American Statistical Association,* 83, 150-159.