SUBSAMPLING CALLBACKS TO REDUCE SURVEY COSTS

Michael R. Elliott, Steven Lewitzky, and Roderick J.A. Little, University of Michigan
University of Michigan, Dept. of Biostatistics, 1420 Washington Heights, Ann Arbor, MI 48109

Key Words: Sample survey, survey costs, subsampling, optimal allocation.

Abstract:

Population surveys with multiple callbacks to maximize response rates are expensive, and a substantial proportion of the data collection cost comes from the many callbacks required to obtain a small proportion of interviews with difficult-to-reach respondents. We explore whether data collection costs are reduced by subsampling a random proportion of the originally sampled units from the mth callback attempt forward, with case weights utilized to adjust for the undersampling, subject to the constraint of maintaining a constant variance of an estimated mean.

1. Introduction

Population surveys often require multiple callbacks to maximize response rates, and a substantial proportion of the data collection cost comes from the many callbacks required to obtain a small proportion of interviews with difficult-to-reach respondents. To reduce costs, a random proportion α of the remaining callbacks are subsampled at the mth callback attempt for continued effort, and the remaining $1-\alpha$ proportion dropped. Case weights are utilized to correct for any bias that might result from the undersampling of more difficult-to-reach respondents. Variance is kept constant by increasing the initial number of sampled units.

The paper shows that subsampling can be cost-effective whenever the proportion of total cost associated with the remaining callbacks exceeds the proportion of interviews obtained from the callbacks. That is, when:

1. the probability of obtaining an interview on a given callback is declining and
2. the callback or interviewing costs for these late respondents are greater on average than for the earlier respondents.

1.1 Previous Work

Previous work in this area has been conducted by Deming(1953) and Hansen and Hurwitz(1958). Deming assumed that response probabilities were constant over calls, and made no allowances for the cost of refusals. Hansen and Hurwitz restricted their analysis to the two-callback, mixed mode case. For an excellent summary of these approaches, see Groves(1989). This paper extends these works by allowing for differings response probabilities at each callback, non-zero costs of refusals, and multiple calls up to some fixed number K.

2. Description of Full Callback and Subsampling Strategies

For the full callback (FC) strategy, up to K callbacks are made for all sampled units.

For the subsampling (SS) strategy, at the mth callback attempt ($m < K$), a random proportion of α nonresponding units are retained for future callback attempts; all attempts are terminated for the remaining $(1-\alpha)$ proportion. Up to K callbacks are attempted for the proportion retained in the sample.

We presume to know or estimate:

p_j = P(interview at the jth callback)

q_j = P(interview at the jth callback | no interview or refusal up to the jth callback)

r_j = P(refusal at the jth callback)

s_j = P(refusal at the jth callback | no interview or refusal up to the jth callback)

c_j = cost of the jth callback

d_j = cost of an interview at the jth callback

e_j = cost of a refusal at the jth callback

N = number of units initially in the sample under FC strategy

N' = number of units initially in the sample under SS strategy

2.1 Expected Costs and Variance of Mean Under FC and SS Strategies

To determine an *efficiency ratio*, find the expected total cost under the two strategies as NE (FC) and $N'E_\alpha$ (SS), where E and E_α are the expected per-

This research was supported by the National Institute of Mental Health grant MH 46376 and MH 49098, with supplemental support from the National Institute of Drug Abuse and the W.T. Grant Foundation. We also thank Ronald Kessler, PhD., Harvard University, for additional comments.

unit cost under the FC and SS strategies, respectively.

We then determine N' as a function of N by requiring that the variances of a sample mean be equal under the two strategies.

We assume for simplicity that our survey sample is drawn using simple random sampling. Then stratifying by callback and conditioning on the callback strata sample sizes,

$$Var(\overline{x}) = \frac{\sum_{j=1}^{K} p_j \sigma_j^2}{N[\sum_{j=1}^{K} p_j]^2}$$

where σ_j^2 is the variance of element x in callback stratum j. Weighting by $1/\alpha$ to correct for the underrepresentation of respondents requiring m of more callbacks,

$$Var_\alpha(\overline{x}) = \frac{\sum_{j=1}^{m-1} p_j \sigma_j^2 + (1/\alpha)\sum_{j=m}^{K} p_j \sigma_j^2}{N'[\sum_{j=1}^{K} p_j]^2}$$

Thus $Var(\overline{x}) = Var_\alpha(\overline{x})$ implies

$$N' = \frac{N(\sum_{j=1}^{m-1} p_j \sigma_j^2 + (1/\alpha)\sum_{j=m}^{K} p_j \sigma_j^2)}{(\sum_{j=1}^{K} p_j \sigma_j^2)}$$

Also, we have for the expected per-unit costs under the full callback and subsampling strategies:

$$E = \sum_{j=1}^{K} p_j[\frac{c_j}{q_j} + d_j] + r_j e_j$$

$$E_\alpha = \sum_{j=1}^{m-1} p_j[\frac{c_j}{q_j} + d_j] + r_j e_j + \alpha(\sum_{j=m}^{K} p_j[\frac{c_j}{q_j} + d_j] + r_j e_j)$$

and $E_\alpha = E - (1-\alpha)\sum_{j=m}^{K} p_j[\frac{c_j}{q_j} + d_j] + r_j e_j$.

2.2 When is the Subsampling Strategy More Efficient?

For the total cost under the subsample strategy to be less than the total cost under the "full callback" strategy, we need

$$N'E_\alpha < NE$$

or

$$\frac{N(\sum_{j=1}^{m-1} p_j \sigma_j^2 + (1/\alpha)\sum_{j=m}^{K} p_j \sigma_j^2)}{(\sum_{j=1}^{K} p_j \sigma_j^2)} \times$$

(E - (1-α) $\sum_{j=m}^{K} p_j[\frac{c_j}{q_j} + d_j] + r_j e_j) < NE$

Dividing both sides of the above inequality by NE yields an *efficiency ratio*

$$f = \frac{(A' + (1/\alpha)(A - A')}{A}(1 - ((1-\alpha)D/E))$$

where
$A = \sum_{j=1}^{K} p_j \sigma_j^2 \qquad A' = \sum_{j=1}^{m-1} p_j \sigma_j^2$
$D = \sum_{j=m}^{K} p_j[\frac{c_j}{q_j} + d_j] + r_j e_j$
$E = \sum_{j=1}^{K} p_j[\frac{c_j}{q_j} + d_j] + r_j e_j$

Assume the variance of our measure of interest is constant across the callback strata; then f reduces to

$$f = (c + (1/\alpha)(1-c))(1 - (1-\alpha)c'))$$

where $B' = cB$ and $D = c'E$ for $B = \sum_{j=1}^{K} p_j$ and $B' = \sum_{j=1}^{m-1} p_j$. Note that c is the probability of an interview occuring before subsampling given that an interview occurs and c' is the proportion of expected cost incurred after subsampling.

To minimize f with respect to α and m, fix m and thus c and c'. Differentiating f with respect to α, setting equal to zero, and solving for α yields

$$\alpha = \sqrt{\frac{(c'-1)(c-1)}{cc'}}$$

In order for α and thus f to be less than 1, $c + c' \geq 1$ – that is, when the proportion of the cost in the callbacks to be subsampled (c') is greater than the proportion of interviews remaining $(1-c)$, subsampling will be the more efficient strategy.

	$c' = 0.25$	$c' = 0.50$	$c' = 0.75$	$c' = 0.90$
$c = 0.25$			$\alpha = 1.00$ f=1.00	$\alpha = 0.58$ f=0.96
$c = 0.50$		$\alpha = 1.00$ f=1.00	$\alpha = 0.58$ f=0.94	$\alpha = 0.33$ f=0.80
$c = 0.75$	$\alpha = 1.00$ f=1.00	$\alpha = 0.58$ f=0.94	$\alpha = 0.33$ f=0.75	$\alpha = 0.19$ f=0.56
$c = 0.90$	$\alpha = 0.58$ f=0.96	$\alpha = 0.33$ f=0.80	$\alpha = 0.19$ f=0.56	$\alpha = 0.11$ f=0.36
$c = 0.95$	$\alpha = .40$ f=.91	$\alpha = .22$ f=0.72	$\alpha = .13$ f=0.46	$\alpha = 0.08$ f=0.27

Table 1: Minimum Efficiency Ratio and Minimizing Value of α as a function of c (given an interview occurs, probability of an interview occuring before subsampling) and c' (proportion of expected cost incurred after subsampling callback in the absence of subsampling).

Note that savings increases as the proportion of total cost (under full sampling) from the callbacks to be subsampled increases while the proportion of interviews from these callbacks (and thus the degree to which the weights increase the variance) decreases.

3. Subsampling when Probability of Obtaining Interview and Cost of Interview is Constant and Callback Cost Increases Linearly

(In the remaining sections we assume for ease of presentation that refusal costs are zero.) Assume a) $K = 10$, b) interview cost is fixed, c) conditional probability of interview is constant ($q_j = q$ for all j), and d) callback cost increases linearly ($c_1 = 1, c_2 = 2, \ldots$). Then total savings of 0.5-7% appear achievable, depending on the ratio of the interview and callback costs and on the probability of obtaining an interview at a given callback (see Figure 1).

As the cost of the interview relative to the callback increases, the effectiveness of subsampling diminishes since the callback cost as a proportion of total cost diminishes. As the callback success rate increases, an increasing proportion of interviews are completed during early callbacks, implying a) earlier subsampling, b) declining savings under subsampling.

3.1 Callback Cost Has Large Step Function

If bring forward our assumptions from the paragraphs above but further assume that we have a mixed-mode contact (e.g., postal followed by face-to-face), so that callback costs increase linearly within mode but that the cost of the second mode is 100 times greater for callback/interviews than initial mode, and further that the more expensive mode is used from the 3rd contact attempt forward, then clearly subsampling should begin at the 3rd callback, and should be quite severe. The potential savings is substantial - approaching 45% when the probability of initial interview is 0.33 (see Figure 2).

Here, savings are reduced if q is relatively small, since the increase in sample size required to compensate for the larger proportion of interviews with large weights (e.g., $1/\alpha$=20) will "eat up" some of the savings generated through subsampling. Also, savings achieved is relatively independent of the cost of the the interview relative to that of the callback within each mode - the operative factor is dramatic increase in callback and interviewing costs at the third callback.

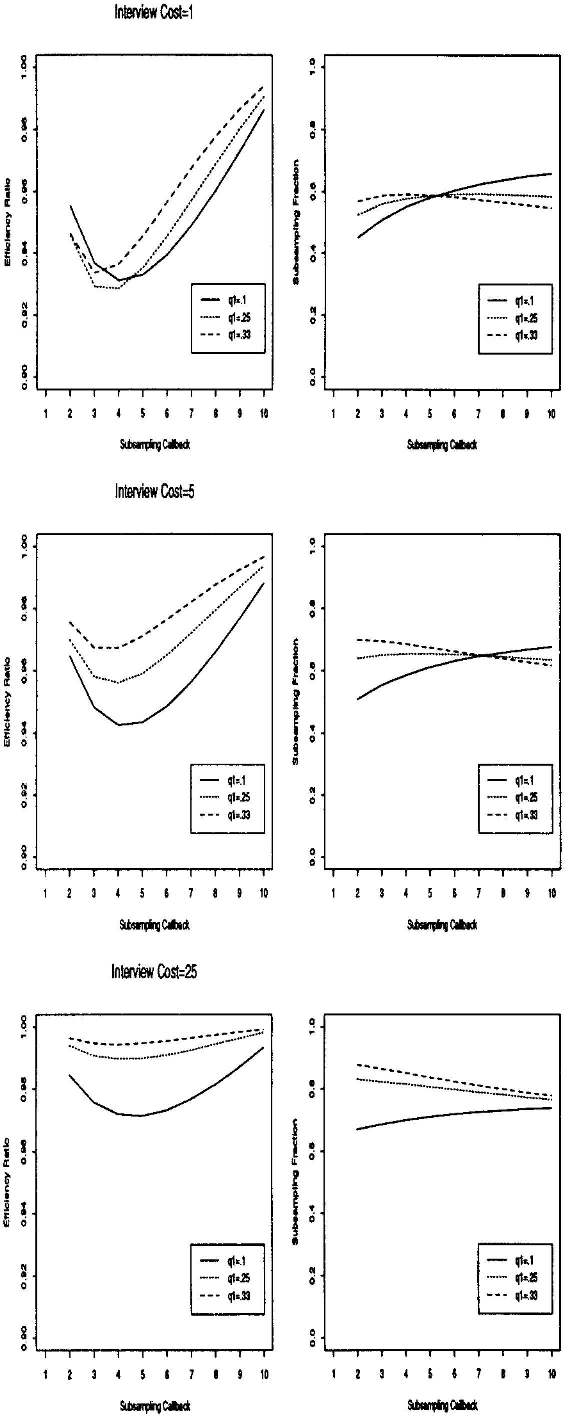

Figure 1: Efficiency ratio and subsampling fraction by callback, where interview cost is fixed at 1, 5, or 25 times of initial callback, and callback cost increases linearly.

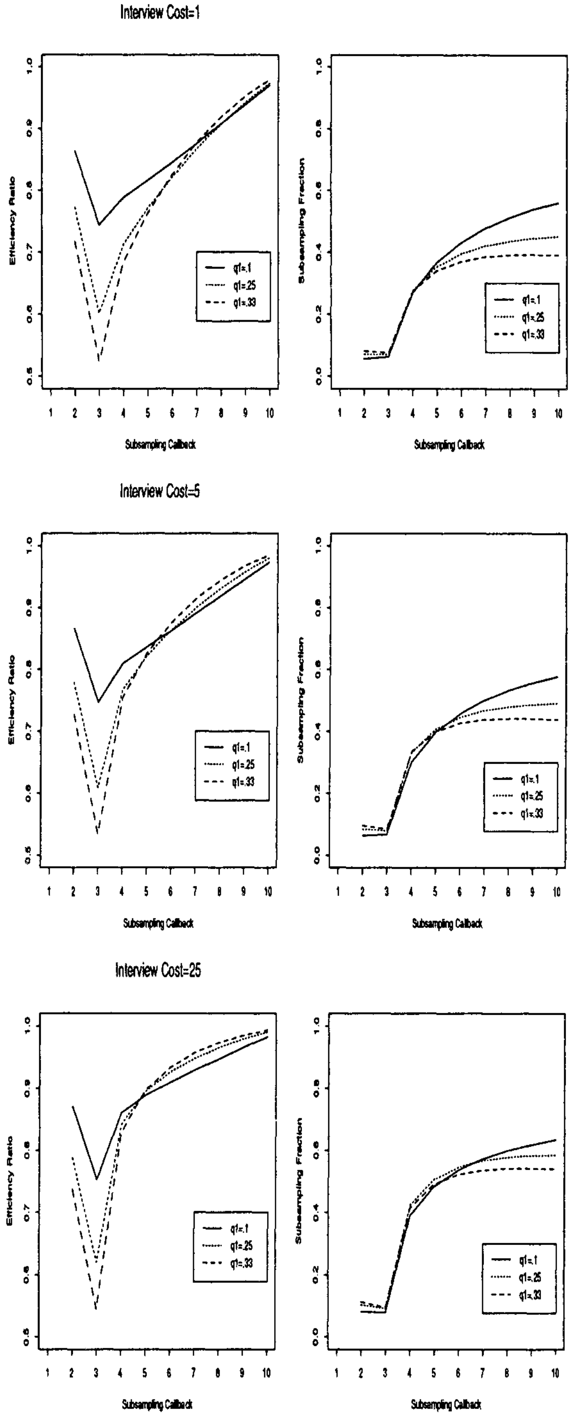

Figure 2: Efficiency ratio and subsampling fraction by callback, where interview cost is fixed at 1, 5, or 25 times of initial callback, and first and second callback/interview costs are 0.01 that of later callback (within mode, callback cost increases linearly).

4. Relationship to Optimal (Neyman) Allocation for Stratified Sampling

Neyman (or optimal allocation) procedure for stratified sampling (Kish, 1965) says that the total variance of the sample will be minimized if

$$n_h \propto W_h S_h / \sqrt{(}J_h)$$

where n_h is the sample from strata h, W_h is the proportion of the population in strata h, S_h is the standard deviation of an obervation in stratum h, and J_h is the average cost of an interview within stratum h.

Consider a two callback case, where the callback cost is essentially 0 but the cost of the interview is D times greater in the first callback than the second. Then, assuming the standard deviation is constant, we obtain the following:

Callback	No Subsampling	Subsampling	Neyman Allocation
First	q_1N	q_1N	q_1N
Second	$q_1(1-q_2)$N	$\alpha q_1(1-q_2)$N	$\frac{q_1(1-q_2)N}{\sqrt{D}}$

Table 2: Expected sample sizes under full callback, subsampling, and optimal allocation

Given that the "sunk costs" of the initial interview are 0 and that we condition on the "observed" sample, our optimal subsampling for callbacks converges to Neymann allocation. In particular, applying our previous analysis to this scenario yields $\alpha = 1/\sqrt{D}$.

5. The Efficiency Ratio Under the Assumption of Heteroscedacity

If variances are not considered to be constant across the strata, the efficiency ratio becomes

$$f(k, c', \alpha) = (k + (1/\alpha)(1-k))(1-(1-\alpha)c'))$$

where $A' = kA$ and $D = c$ for

$A = \sum_{j=1}^{K} p_j \sigma_j^2 \qquad A' = \sum_{j=1}^{m-1} p_j \sigma_j^2$

$D = \sum_{j=m}^{K} p_j [\frac{c_j}{q_j} + d_j] + r_j e_j$

$E = \sum_{j=1}^{K} p_j [\frac{c_j}{q_j} + d_j] + r_j e_j$

This expression is similar to previous results, except that proportion of interviews is now weighted by variance within strata.

Since $1/\alpha \geq 1$, f will be small if the variance of the mth and later callbacks is much smaller than for the earlier callbacks. For example, under the scenario of constant probability of obtaining an interview and increasing costs of callbacks (K=10, q_j=0.25 for all

j, $c_j = j$, d=5), if we assume that the variance in the last five callback strata is 0.5 that of the previous strata,

- subampling should begin at m=6 instead of m=4
- the proportion decreases from α=65% to 46%, and
- Estimated savings increases from 4% to 9%.

6. Application: The National Co-Morbidity Survey

We apply our analysis to the National Comorbidy Survey (NCS, Kessler, 1992) a 1990-92 nationwide face-to-face survey of 8098 US-wide respondents aged 15-54 regarding prevalence of psychiatric comorbidity.

48,258 callbacks were attempted in 16,263 interviewer trips, an average of 3.0 callbacks/trip. All but two interviews conducted by 19 callbacks.

Unfortunately, no cost data are directly available, only data on

- number of trips
- number of callbacks and interviews on a particular trip,
- number of callbacks required to conduct a given interview.

"Guesstimating" relative overhead costs of a trip (getting into a car and driving back and forth to and within a segment, etc.), cost of a single callback attempt having arrived at segement, and additional cost of an interview on a callback attempt allowed us to derive the average cost of mth callback.

If this survey were to be repeated with a maximum K=20 callbacks and similar cost structure, would subsampling be a sensible strategy? Figure 3 suggests the potential savings are modest.

Under a variety of cost assumptions, the greatest savings predicted is only 1%. Maximum savings are obtained by starting subsampling at m=6, with 77-85% of active blocks retained.

Since variability of the estimates ($\propto p(1-p)$) is related to the point estimates themselves, the assumption of homoscedacity is violated. But prevalence of disorders tended to increase with later callbacks, which would lead to an increase in variance among the later callbacks and thus further decrease in the savings estimated above.

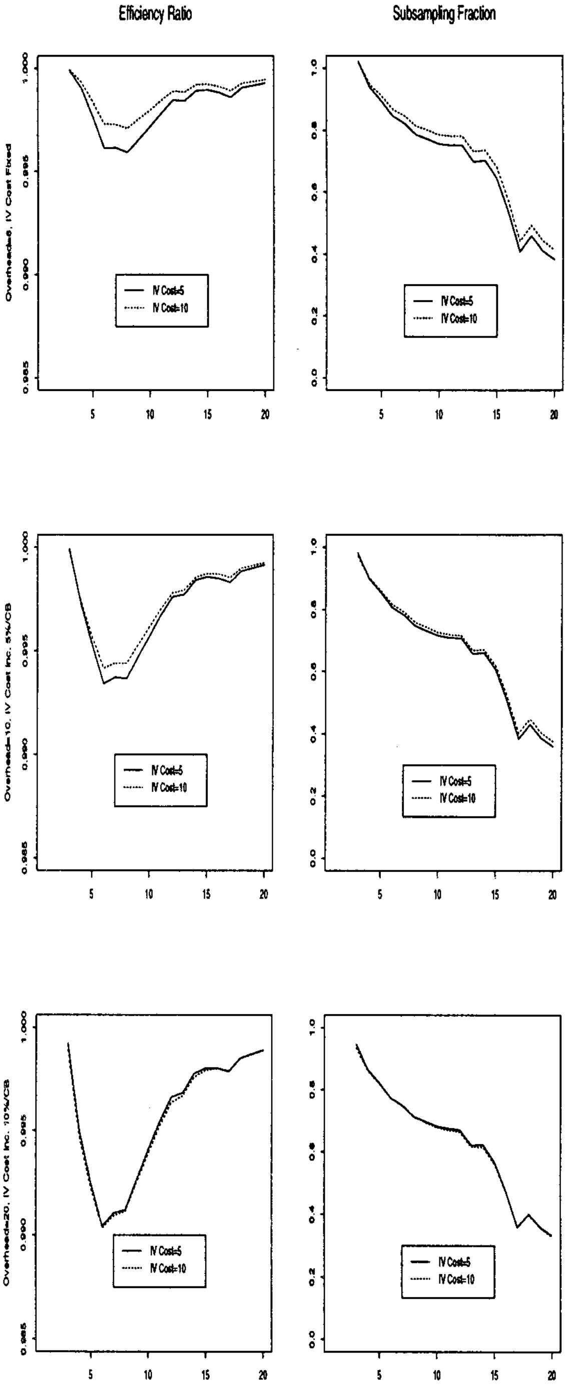

Figure 3: Efficiency ratio and subsampling fraction by callback for NCS, where overhead cost is estimated to be 6, 10, and 20 times than of callback cost, interview cost is fixed or increase at 5% or 10% per callback and is considered equal to 5, 10, or 20 times that of initial callback.

7. Discussion

Subsampling saves resources whenever the proportion of the cost related to callbacks to be subsampled is greater than the (variance-weighted) proportion of interviews to be obtained from these callbacks. In practice, it appears that both of these conditions must be met to a significant degree for substantial savings to occur: a rough "rule of thumb" appears to be that the conditional probability of interview must be declining at least as fast, and the callback cost increasing at least as fast, as the square root of the callback number. More promising is a situation where callback cost has a large step function, as in a postal mailing followed by face-to-face followup with non-respondents (e.g., the U.S. Census).

Large, fixed interview costs will tend to negate the saving achieved through subsampling. Similarly, increased variance of the variable of interest among late callbacks reduces the effectiveness of subsampling, while reduced variance increases its effectiveness.

Application of our results to the 1991 CoMorbidity Survey showed that, for a large face-to-face survey, appropriate subsampling would on average yield savings, although probably not large enough to justify carrying out the procedure. Before implementing a subsampling scheme, it would be wise test the sensitivity of the subsampling fraction and cutpoint to differing cost and variance assumptions.

8. Further Directions

More complex subsampling strategies might include

- Randomly drop a constant proportion of sampled units from the mth callback on,

- Randomly drop a varying proportion of sampled units at each callback as a function of the cost and variance parameters.

Also, further work is needed to determine the effect of more complex (and realistic) survey designs. Clearly, one effect of a multiple-stage cluster sample design would be to require the dropping of segments rather than sampled dwelling units. (If only a portion of dwelling units were dropped within a segment, little savings would be achieved since most of the extra cost associated with later callbacks involves the inefficiency of attempting to contact relatively fewer units in a large geographic area.)

These results focus on *expected* savings that would result under the subsampling strategy; no attempt havs been made to quantify the variability associated with these expected values.

Finally, we note that a model-based rather than a design-based approach to analysis may yield increased efficiency from subsampling schemes. Rather than fully correct for bias by inflating the weights of late callback by $1/\alpha$, a modeling strategy focused on minimizing mean square error might attenuate weighting effects and hence lead to improved efficencies for subsampling. Also, for statistics other than means and totals (e.g., regression coefficients), the variance inflation from weighting subsampled cases may be reduced, thus making subsampling a potentially more attractive strategy.

9. References

Deming, WE. "On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Response and the Bias of Nonresponse," *Journal of the American Statistical Association*, 48:264,743-772,1953.

Hansen, MH; Hurwitz, WN. "The Problem of Nonresponse in Sample Surveys," *Journal of the American Statistical Association*,517-529,1958.

Groves, RM. *Survey Errors and Survey Costs.* New York: John Wiley and Sons, 1989.

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

Kessler, R. National Comorbidity Survey. Ann Arbor, MI: Survey Research Center, Institute for Social Research, 1992.