*The Canadian Journal of Statistics*
Vol. 16, No. 2, 1988, Pages 135–151
*La Revue Canadienne de Statistique*



# From the inside out and the outside in: Combining experimental and sampling structures

Stephen E. FIENBERG and Judith M. TANUR

*Carnegie Mellon University* and *State University of New York at Stony Brook*

*Key words and phrases*: External validity, internal validity, interviewer effects, randomized experiments, sample surveys.

*AMS 1985 subject classifications*: Primary 62D05, 62K10; Secondary 62F99, 62K99, 92A20, 92A25.

## ABSTRACT

The three basic tenets of experimental design (randomization, replication, local control) find parallels in sampling design. While the ways these parallel structures are applied differ across the two areas, the commonalities suggest ways of strengthening work in both. We describe examples of embedding sampling within experiments, the use of experimental design structures within sample surveys (including interpenetrating networks of samples), and the generalization from experiments to populations through random selection and sampling. A more complicated connecting structure between experiments and surveys is illustrated by the U.S. Census Bureau model for nonsampling errors. We conclude with a description of alternative approaches to basic inference in complicated embedded structures.

## RÉSUMÉ

Les trois principes de base en schémas expérimentaux (randomisation, réplique, contrôle local) ont leur parallèle en planification de sondages. Quoique ces structures parallèles soient appliquées de façon différente dans les deux domaines, leurs aspects communs peuvent être la source de travaux approfondis dans un domaine comme dans l'autre. Nous décrivons des exemples d'échantillonnages enchâssés dans des plans d'expérience, l'utilisation de schémas expérimentaux à l'intérieur de sondages (incluant les réseaux d'échantillons s'interpénétrant) ainsi que l'extrapolation des résultats d'une expérience à la population grâce aux techniques d'échantillonnage. Une structure connexe entre expérience et sondage, mais offrant plus de complexité, est illustrée par le modèle du U.S. Census Bureau pour erreurs autres que d'échantillonnage. Nous concluons en décrivant de nouvelles approches dans le traitement des structures enchâssées complexes et qui sont des solutions de rechange à l'inférence fondamentale.

*This research was supported in part by the National Science Foundation under Grants No. SES-84-06952 to Carnegie Mellon University and Grant No. SES-84-06721 to the Research Foundation of the State University of New York. We are indebted to the following who provided us with examples, insights, recollections, and references: Barbara Bailar, Paul Biemer, David Brillinger, James Durbin, Ivan Felligi, David Finney, Robert Groves, Daniel Horvitz, Oscar Kempthorne, Richard Link, Barry Margolin, Howard Schuman, Chris Scott, Benjamin Tepping, Roger Tourangeau, Joseph Waksberg, and Andrew White. Two referees made helpful comments that sharpened the focus of the paper.

## 1. INTRODUCTION

Random sampling and randomized experimentation are inextricably linked. Beginning with their common origins in the work of Fisher and Neyman from the 1920s and the 1930s, one can trace the development of parallel concepts and structures in the two areas (see Fienberg and Tanur 1985, 1987). One of the more important lessons to be learned from the parallel concepts and structures is that they can profitably be linked and intertwined, with sampling embedded in experiments and formal experimental structures embedded in sampling designs. In this paper, we review some of the ways that experimental and sampling structures have been combined in statistical practice and the principles that underlie their combination; we also make some suggestions towards the improvement of practice.

The three basic tenets of experimental design as advocated by Fisher (1935) are randomization, replication, and local control. These are paralleled almost perfectly by concepts in sampling design. There are, of course, substantial differences in the way these parallel structures are used in the two areas [e.g. see the discussions in Fienberg and Tanur (1985, 1987)], but the commonalities are sufficiently great to suggest how links can be forged to strengthen work in both areas. Take the principle of local control, which has the experimental purpose to ensure, through the homogeneity of experimental units, that actual differences will be detected with high probability. Thus, if an experiment is to be embedded in a sample design, the principle of local control suggests that the embedding be done in a way that takes greatest advantage of the homogeneity of observations on sampling units. In sample surveys one of the greatest sources of observational variability is the nonsampling error due to interviewer differences [e.g. see Bailey, Moore, and Bailar (1978)]. When we are not planning to measure interviewer differences explicitly, then we can exercise local control by embedding replications of the experiment within interviewer. (In such a design each interviewer is treated as a block.) These forms of embedding have been used successfully and are described in greater detail in Section 3.

We can discern five main purposes for embedding sampling or sample surveys in experiments and vice versa:

(i) to sample *data* or *treatments* within experiments;

(ii) to broaden the scope and extend the generalizability of experiments through sampling;

(iii) to compare alternative aspects of survey methodology (questionnaires, training methods, collection methods) either in pilot surveys, in methods test panels, or in ongoing surveys;

(iv) to make comparisons of substantive (rather than solely methodological) interest;

(v) to explore the components of response variation and the validity of surveys.

In Section 2, we illustrate (i) and (ii) by describing some early examples of embedding, the ways in which statisticians originally described the linking of sampling and experimentation, and some more recent examples of sampling embedded in experiments. Then, in Section 3, we consider (iii), (iv), and (v), and elaborate upon the use of experimental design structures within sample surveys, including the conception of interpenetrating networks of samples and the voluminous work of statisticians at Statistics Canada, the U.S. Bureau of the Census, and elsewhere. In

most of the illustrations of embedding, the structure of either the sample design or the experimental design may be complex, but the connecting structure is usually relatively simple in form. In Section 4, we describe the U.S. Census Bureau model for nonsampling errors and discuss experimental-design aspects of the measurement of its components. In Section 5, we discuss the issue of generalization from experiments to populations through random selection and sampling. The issue of embedding experiments within surveys raises questions of statistical inference that are rarely discussed. In Section 6, we describe some alternative approaches to basic inferences in such problems.

## 2. EMBEDDING SAMPLING WITHIN EXPERIMENTS

### *2.1 Some Early Examples*

The agricultural work at Rothamsted in the 1920s and 1930s, at the time that basic ideas of sampling and experimentation were being developed, combined ideas from both areas explicitly. Although the focus of the work was experimentation, frequently it was neither necessary nor feasible to measure yield on the entire experimental unit or plot. Thus the idea of sampling crops to obtain observations of yield within plot was introduced at Rothamsted in 1920 (see e.g. Clapham 1931). Yates (1985) attributes the formal use of analysis-of-variance breakdowns for such sampled data within experiments to Fisher, but others at the time were using related ideas (see Robinson 1987).

For example, Yates and Zacopany (1935) consider a problem involving the sampling of units within plots. The structure permits them to model three sources of variation: within-plot error variance, unit-to-unit sampling variance, and a component due to competition between sampling units. On top of the experimental design that assigns treatments, Yates and Zacopany impose a sampling design to measure the outcomes of the treatments. They go on to derive the efficiency of sampling compared with complete harvesting, to develop formulae for optimal amounts of sampling in terms of costs (or work involved), and to apply their methods to a series of 18 experiments involving cereal crops at Rothamsted.

Yates (1935) described a variety of biases that might arise in the sampling of agricultural materials where the observer's choice has influenced the selection of the sample, and Cochran and Watson (1936) report on an experiment designed to measure that sampling bias:

> [S]ix sampling-units were marked out, and the height of every shoot in each was measured. The observers were told to measure the height of two shoots in each row as usual, but instead of picking the end shoots, they were to select any two in the row which they pleased, so as to give what they considered to be a random sample of the shoot-heights in the sampling-units. There were twelve observers, who were divided into two groups of six. The members of the same group took their observations at the same time, but the order in which they made their selections from the various sampling-units was fixed by means of a $6 \times 6$ Latin-square arrangement, so that no two observers were measuring the same sampling-unit together. There were above 20 shoots per row within a sampling-unit, and the experiment was made about a week before ear-emergence, the wheat being 70 cm. high. (p. 70)

There are two levels of embedding here, since this is an experiment to study a sampling issue relevant to the analysis of an experimental design. Moreover, since

there are a pair of observations per row, Cochran and Watson were able to get a "proper" estimate of the sampling variation within observers (in a sense a third level of embedding). After an extended analysis of the experimental results the authors concluded:

> This experiment, in short, very strongly supports the evidence from other investigations that the only sure method of avoiding bias is for the sampling to be a random. (p. 75)

### *2.2 More Recent Examples: Sampling Treatments and Control Structures.*

In addition to the sampling of material within plots of an experiment, other forms of sampling are used either explicitly or, more often, implicitly in the sampling of levels of treatments (e.g. see Wilk and Kempthorne 1956) and in the sampling of blocks, rows, and columns, or more complex-structured forms of control (e.g. see Federer 1976a, 1976b, 1977). For example, in the context of Latin-square designs, Federer (1976b) describes two sampling structures that are alternatives to the traditional notion of dividing a piece of land up into $r$ rows and $c$ columns. He begins with a population divided into $R$ rows and $C$ columns and then:

(1) obtains a simple random sample of r rows and a simple random sample of $c$ columns, and then randomly selects one experimental unit (*e.u.*) from each of the $rc$ subpopulations, or

(2) conceives of units of size $r$ *e.u.*'s by $c$ *e.u.*'s, and then conceives of these units of $r \times c$ *e.u.*'s as coming from a population composed of such groupings into units from the $R$ rows and $C$ columns. Then the experimenter chooses one unit at random from the population and lays out an $r$-row by $c$-column design on the selected unit.

Federer suggests standard additive (or multiplicative) ANOVA models with independently and identically distributed error terms to go with such sampling schemes.

Rossi and his colleagues (e.g. Rossi and Anderson 1982) have used a design that samples treatment levels in order to embed a large factorial experiment within a sample survey. Typically respondents are asked to render judgments about social objects that can be described on $K$ dimensions with $q_k$ levels for dimension $k$. Thus the object universe consists of $\Pi_k q_k$ objects—which can be a very large number if $K$ and several of the $q_k$ are moderately large. The strategy is to construct an "object sample" for each respondent to rate by constructing objects through simple random sampling from the levels of each dimension. Each respondent has the task of rating a different random sample of objects from the object universe; the ratings from these random sample are combined over respondents, so that estimates of the effects of the dimensions (and their interactions) are combinations of between-respondent and within-respondent effects. This approach neither guarantees the estimability of all main effects and interactions, nor offers as much opportunity for the exercise of local control as would the use of a collection of fractional factorials, perhaps using a BIBD-like variant of Federer's approach of subsampling levels within dimensions, possibly separately for each respondent.

## 3. EMBEDDING EXPERIMENTS WITHIN SURVEYS

Marketing surveys attempting to manipulate several factors expected to influence consumer preference occasionally take advantage of elegant experimental designs in order to gain maximum information from each respondent. For example, A.J.

Wood Research Corporation (Wood 1959) described a plan to use Latin-square structures to construct carefully balanced possible consumer-choice combinations in order to determine the effects of type of store, brand, and distance from the consumer's home on preferences for brands of ice cream. Most survey uses of experiments are less elaborate: the simplest is called a split-ballot (although it should be more accurately called a split-sample) approach to experimentation.

### *3.1 Split-Ballot Techniques and Alternatives.*

Traditional split-ballot experimenters take two (or more) versions of a questionnaire and administer each to a fraction of the sample—or, to be more precise, to two (or more) independent but similarly structured samples. Investigators usually make no formal attempt to interlock the sample design and the experimental design, and typically they compare the similarly structured samples directly, ignoring whatever interlocking sampling features are in place. For example, if the same clusters are used for two or more questionnaires, this interlocking feature needs to be built into the analysis.

Schuman, Steeh, and Bobo (1985) investigating racial attitudes in America, conducted a split-ballot experiment in January 1983 in which half of a national telephone sample were asked the general desegregation item after the federal school intervention item and the other half were asked the questions in reverse order. They found that the percentage of respondents endorsing desegregation dropped from 61.4% to 38.9% when the general desegregation question was preceeded by the item on federal school intervention. In the tradition of split-ballot experimentation, the authors neither describe how the two subsamples are structured, nor do they use anything in the structure of the subsamples as part of their analyses. Such experiments are often carried out within ongoing surveys and must take the survey design as given. This approach should be contrasted with the explicit design of a survey to facilitate experimental comparisons, illustrated in the following example.

To address methodological questions concerning the effects of questionnaire context on responses to attitude items, investigators at NORC used issues at three differing levels of familiarity (a within-respondent factor), and manipulated context, content (positive or negative), and depth of thought (by presenting an open-ended probe of the respondents' thought processes early in the questionnaire or at the end). Cases were selected as a SRS from telephone banks listed in the Chicago directory, and each interviewer's assignment consisted of several replications of the $3 \times 2 \times 2 \times 2$ experimental design (Tourangeau 1986). Thus interviewers were used as blocks. Even modest interviewer (block) effects can be important here, because the efficiency of blocking (e.g., see Cochran and Cox 1957, p.112) increases both with the block size and with the number of blocks. We note, however, that this design has the possible drawback that the levels of the blocking variable are human beings, the interviewers. Such interviewers may well change their behavior as they administer differing forms of the questionnaire, thus creating artifactual effects similar to the experimenter-expectation effects described by Rosenthal and his colleagues (Rosenthal and Rubin 1979).

### *3.2 Interpenetrating Networks of Samples.*

A classic instance of the embedding of an experimental structure within a sampling framework, due to Mahalanobis (1946), is the method of *interpenetrating*

*networks of samples* (IPNS), which provides a built-in replication structure for validating survey results (see also Bailar 1983). For example, in a survey on the economic conditions of factory workers in an industrial area of India, Mahalanobis divided the area into subareas, and arranged for the selection of 5 independent random samples within each subarea. Each of 5 interviewers worked in all subareas. This IPNS design thus provided 5 independent estimates of the economic conditions, and as a consequence allowed for an evaluation of the response variation associated with interviewers (see also Hansen et al. 1953b, Chapter 12). In the absence of interviewer effects, an IPNS design gives an internal estimate of variability without direct reference to the probability aspects of a complex sample design—a precursor to the modern literature on replication and jackknifing for variance estimation in surveys (see Kish and Frankel 1974). Note, however, that there is a tension here—the internal estimate of variability (i.e., sampling error) will be confounded with interviewer variability unless interviewer effects are either assumed absent or estimated separately from another level of replication in the design.

A subsequent literature on interviewer variance has gone in a variety of directions. Kish (1962), for example, examined a pair of studies where the direct measurement of the effects of interviewers was feasible because of the absence of complex survey sample structure. He studied the intraclass correlation resulting from interviewer variance and then considered the optimum number of interviews per interviewer, based on cost factors.

Yates (1981, pp.110–111) describes a different approach which makes possible the separate estimation of interviewer effects and a measure of internal variability by using *local control within interviewer* to compare alternative questionnaires together with the *measurement of interviewer differences* through an IPNS-like structure. Combinations of questionnaire form and interviewer are randomly assigned within blocks of respondents in a $2 \times 3$ factorial design in randomized blocks. With this example as a starting point, one can visualize other examples of relatively complex embeddings of IPNS structures to achieve useful variations on traditional experimental designs.

The original IPNS idea in which the sample is broken up into fully replicated subsamples represents an ideal case in which the costs of interviewer travel to reach sample units widely dispersed over the population is negligible or at least affordable. But in reality financial and human cost factors combine to render interviewers much less mobile than the IPNS ideal assumes, and although ambitious travel plans can be undertaken occasionally, more usually compromise designs involving restricted randomization must be sought. For example, Fellegi (1964) combined partial IPNS and reenumeration to estimate the components of a model of response error in connection with the 1961 Canadian Census of Population. A pair of contiguous enumeration areas (EAs) was sampled from each of 67 strata. A pair of enumerators was assigned to each stratum and a sample of addresses assigned at random to each enumerator. On reenumeration these samples were interchanged within each stratum. Thus, although the design has enumerators nested within strata (instead of crossed with strata as in a full IPNS design), this combination of partial IPNS and reenumeration permits the estimation of more of the parameters in the responses error model than would be possible with either method alone.

The advent of telephone interviewing and its widespread application in many large-scale surveys present the opportunity to return to the original conception of

IPNS. Long-distance telephone charges remain the same regardless of whether one or several interviewers are placing the calls to a single area code. One can begin with a large sample of telephone numbers grouped according to 3-digit exchanges or banks of numbers. Then this large sample can be broken into interpenetrating subsamples, and each subsample assigned to an interviewer. In this way problematic banks of numbers are spread across interviews and are not confounded with productivity differences among interviewers. Implementing such interpenetrating designs can prove difficult when telephone interviewers work in shifts, and thus Stokes (1986) describes an IPNS variant with interpenetrated assignments only within shifts (see also Groves and Magilavy, 1986).

### *3.3 Blocking on Interviewers and Clusters.*

When variations in interview procedure are being investigated, the principle of local control suggests that blocking on interviewer is appropriate, with each interviewer using several or all of the varying procedures. This description is similar to that of a split-plot experiment with "blocks" corresponding to the grouping of subplot units into whole plots. This, however, is not quite our intent. In Fienberg and Tanur (1985, 1987) we note that different levels of clustering in a sampling plan correspond to different levels of plots in a split-plot design. Thus, at each level of the plan one can incorporate an appropriate design, possibly with forms of blocking and treatment structure (e.g. see Federer 1977). A cluster of households assigned to an interviewer in a household survey thus corresponds to the lowest level of a split-plot experimental unit. In this sense clusters are confounded with interviewers. If the size of this lowest-level cluster is sufficiently large (as it may be in a telephone survey), then an additional level of blocking (or stratification) can be used within interviewer for even more precise comparisons. On the other hand, when an interviewer is assigned several clusters of households, interviewers correspond to blocks at a whole-plot or intermediate-plot level, and if treatments are assigned at the level of interviewers, only interviewer-by-treatment interactions can be examined at the subplot level. To understand how to analyze such experiments embedded within surveys, the statistician needs a good working knowledge of the analysis of nontrivial split-plot experiments.

It is also important to note that in many surveys, especially those employing variants of area sample (see Kish 1965, pp. 301–358), there is substantial variation among clusters or geographical segments relative to variation within. Since it is often economical to employ a single interviewer within a cluster or segment, much of the gain due to blocking on interviewer may really be attributable to segments. Nonetheless, for simplicity we continue to focus on interviewers as the locus of control.

When one of us suggested blocking on interviewers several years ago at a meeting on sample surveys, someone in the audience commented that giving an interviewer two or more forms of questionnaires to administer risked confusion and would result in useless responses. Confusion would be minimized, according to this argument, if the questionnaires were given to different but parallel samples with different interviewers. This concern, that blocking on interviewers is inadvisable because it is too difficult to carry out, was addressed earlier by Durbin and Stuart (1951), who designed a $3^3 \times 4 \times 2$ factorial experiment completely crossing three survey organizations, three types of questionnaires, three interview areas in London, four ages of respondents, and two sexes. Further, within one of the survey organiza-

tions they completely crossed age of interviewer and sex of interviewer. Each interviewer, while confined to only one district, handled all three questionnaires in approximately equal numbers with an approximate balance of age and sex groups of respondents. The finding of this study was that inexperienced student interviewers had statistically significantly lower response rates than did experienced interviewers. Commenting on the purported difficulty of carrying out such investigations, Durbin and Stuart (1951) remark (p. 184):

> Although highly elaborate designs are often used in other sciences, it is not unnatural that in a field in which the experimental material is composed of human beings, the tendency should have been towards simplicity of layout. In our own experience, however, the extra amount of organization necessitated by the design we used proved to be a good deal less troublesome than had been expected.

This lesson seems to have been only partly assimilated into practice by the U.S. Bureau of the Census in its 1976–77 mode-of-interviewing experiment for the National Crime Survey (NCS). Interviewers were indeed crossed with treatments (usual NCS procedure as a control, experimentally maximizing in-person interviewing, and experimentally maximizing telephone interviewing), but Woltman, Turner, and Bushery (1980) report no control for within-interviewer variability to improve the precision of the reported results. Further, the Census Bureau assigned segments (clusters of housing units with expected size 4) to treatments rather than randomizing the treatments within segments. In support for this design, the authors cited cost efficiency and noted "that erroneous application of treatments could have resulted more often because the units designated to receive the experimental treatment could have been easily overlooked by the interviewer" (p.535). Thus they secured some insurance against interviewer error at what may have been a high cost in sampling error and the confounding of mode of interview effects with segment effects.

Interviewer training, preparation of questionnaire packets in prearranged order, and supervision must be very careful if these experimental strategies of blocking on interviewers are to be used, but such care should pay off richly in increased precision of estimates. Indeed, there is a strong oral tradition (lacking, however, extensive surviving written documentation) that blocking on interviewers was frequently done in the Census Bureau's methodological studies in the 1940s and 1950s. Somewhat more recently, Waksberg and Pearl (1965) describe a methods Test conducted in 1963–64 in which "interviewers in each area were divided into two groups with each group testing two alternative procedures against the standard one used in the Current Population Survey. (It was felt inadvisable to train each interviewer on all of the procedures to be tested.)" Yet, of the 15 comparison tests with surviving documentation conducted by the Census Bureau from 1957 through 1969, this was the only one which blocked on interviewers (see Jabine and Rothwell 1970). Nonetheless, a later study carried out by the Census Bureau for the Committee on National Statistics's Panel on Privacy and Confidentiality as Factors in Survey Response (1979) shows the importance of blocking on interviewers for detecting differences in response rates for different guarantees of confidentiality.

Two additional examples are illustrative. In surveys involving repeated measurements for the same household or respondent, the respondent can be used as the block in a design, with different treatments (e.g. recall periods) being used for different interviews with the respondent. The heuristic link here is that a repeated-measure design is the same as a split-plot design which is parallel to cluster sampling (e.g. see

Fienberg and Tanur 1987). Scott (1973) describes the use of such a design in a household-budget survey in Botswana to determine the optimal length of recall period. In mail surveys, depending on the sizes of the clusters, fairly substantial experiments can be embedded within clusters. For example, Scott (1961) describes a mail survey on radio and television viewing habits in which 5 factors were used in a complete factorial experiment. The survey used 42 sample clusters of size 96, which allowed for a full replicate of a $4 \times 3 \times 2 \times 2 \times 2$ design within each of 42 blocks.

The foregoing discussion may suggest to some that the authors believe that complex experimental designs can be embedded within surveys with ease. We recognize that the day-to-day exigencies of carrying out surveys in the field typically lead to unequal cluster sizes or unequal numbers of observations within interviewers as well as substantial nonresponse. The existence of such complicating factors presents greater methodological challenges to the statistical analyst, but should not be viewed as an argument against carefully planned embedded designs.

## 4. AN ELABORATION OF EMBEDDING: VARIANCE-COMPONENTS MODELS

A broad area of applicability of experimental ideas within surveys is for the modelling and estimation of nonsampling errors using a random-effects ANOVA model. Pioneering work originated in the U.S. Census Bureau (e.g. Hansen, Hurwitz, Marks, and Mauldin 1951; Hansen, Hurwitz, and Bershad 1961) and at Statistics Canada (e.g. Fellegi 1964) and has been much elaborated [see e.g. Cochran (1968) and Stokes (1986)]. The modelling consists of breaking the response variance into components due to interviewers, coders, supervisors, etc., taking into account that errors introduced by any individual are likely to be correlated over his or her interviews.

Mosteller (1978) presents a simple summary of these modelling ideas. Let $Y_{jt}$ be responses at time $t$ for units $j = 1, \ldots, n$ in a sample. If we can think of the survey as conceptually repeatable, then $Y_{jt}$ is a random variable and we can, for example, use $\bar{Y}_t$ to estimate $Z$, a "true" population quantity. Then we can decompose the deviation of $\bar{Y}_t$ from $Z$ into three basic components:

$$\bar{Y}_t - Z = (\bar{Y}_t - \bar{\mu}_s) + (\bar{\mu}_s - \bar{\mu}) + (\bar{\mu} - Z),$$

where $\bar{\mu}_t = \mathcal{E}\bar{Y}_t$ averaging over the hypothetical replications with the same sample, and $\bar{\mu} = \mathcal{E}\bar{\mu}_s$ averaging repeated samplings. $\bar{Y}_t - \bar{\mu}_s$ is random response error, $\bar{\mu}_s - \bar{\mu}$ is sampling error, and $\bar{\mu} - Z$ is bias. The response variance is then rewritten as

$$\mathcal{E}(\bar{Y}_t - \bar{\mu}_s)^2 = \frac{\sigma^2}{n}\{1 + (n - 1)\rho\},$$

where $\sigma^2$ is the variance of $Y_{jt}$ over $t$, and $\rho$ is the correlation of response errors within a sample.

Investigators have elaborated the model in a variety of directions. For example, in an evaluation program to estimate the interviewer component of variation, another interviewer reinterviews the original respondents to get some handle on the correlation between individuals for different interviewers. Multiple individuals per interviewer, in both the original study and the reinterview program, provide correlations within interviewers—the so-called correlated component. A reinterview program not only can estimate the between interviewer and correlated component contribu-

tions to overall variability, but also can consider the impact of different modes of enumeration in light of the response error structure, with the object of reducing the interviewer component by proposing alternative techniques. The sizes of the interviewer component and correlated response error component relative to the overall error (or to sampling variability) led to support of the use of sampling for some characteristics in the U.S. decennial census. (The sampling variability of a 25% or 5% sample of the population was small compared with the variances associated with known sources of response error, especially those attributable to interviewer.) Indeed, between 1950 and 1960 there was a change from interview to self-enumeration in the census because the 1950 results showed the correlated component of interviewer error to be large relative to the other components. This change, while letting interviewer variability go up as each family supplied its own "interviewer," eliminated the correlated response error. Note that the definition of the correlated component can vary across studies. For example, Bailar and Biemer (1984) refer to the definition implicit in the above discussion as "intra interviewer covariance" and separate out from it the covariance common to all interviewers because, for example, they share a working environment, received common training, etc.

What is the design feature of all this? If there are correlations only within interviewers for the errors associated with pairs of individuals, and if the individuals do not overlap (which is the case except in a reinterview survey carried out for evaluation or in a panel study), then there is a direct analog to a classic split-plot experiment with the error structure laid out in Cochran and Cox (1957). In the reinterview evaluation study, because there is an extra observation for each individual (i.e. replication for individuals as well as for interviewers) we have a form of two-way partially balanced split-plot structure. Note, however, that to consider this replication across individuals when reinterviews are separated in time from original interviews is implicitly to assume that individuals remain constant over at least short time periods and that the first interview does not contaminate the second. Relaxing the first of these assumptions introduces yet another component of variance.

This notion of introducing another component of variance to estimate the effect of a particular source of nonsampling error implies that care must be exercised in the design of experiments. Different levels of blocking for local control are crucial.

## 5. GENERALIZING FROM EXPERIMENTS TO POPULATIONS

The other form of embedding apparent in the early agricultural experimentation literature is the use of a number of different sites, in order to obtain average responses applicable across a region or a country. The sampling of experimental sites certainly was not random, but doubtless the intention was for the sites to be "representative" or for "strategic variation". A problem with such series of experiments, whether sampling of sites is at random or not, is the introduction of two new components of variation. The first and largest new component is due to variety $\times$ environment interaction. For full implementation, this variance component is important. The second component, due to the variation in the magnitude of experimental error over the series, is more problematic, and investigators work hard to standardize procedures across sites. Yates and Cochran (1938) noted these difficulties, but attempts to use elaborate series of experiments continued because only from the results of such series can one make recommendations for general agricultural practice.

In the social sciences, a distinction has long been drawn between "internal

validity" and "external validity" of experiments (Campbell 1957; Campbell and Stanley 1963; Cook and Campbell 1979; Aronson, Brewer, and Carlsmith 1985). Internal validity refers to the defensibility of the cause-effect relationship between the treatment and the outcome within the experiment itself. Experimenters contend against threats to internal validity by standardizing the protocols used with the experimental and control groups so that the experiences of the groups differ only in the applied treatment. Even more importantly, they ensure that the groups are the same *a priori* by randomizing between treatment and control. By successfully defending against threats to internal validity, an experimenter can be reasonbly sure that, *in this particular instance*, the treatment caused the effect.

"External validity" means that the treatment (or the conceptual variable that the treatment was designed to operationalize) would cause similar effects in populations other than the one used in the experiment. Traditionally, scientists respond to the challenge of external validity by taking one of two complementary stances. They may argue that, because the processes that they study are sufficiently universal, their choice of subject population is irrelevant. Or they may later attempt to replicate on populations that are chosen to be very different, on dimensions thought to be relevant to the issue at hand, from the population on which the results were initially established.

Fisher (1935, Section 30) describes a variation on this latter technique in which strategic (but nonrandom) variation in the experimental material is introduced into the initial design of an experiment carried out in Minnesota in 1930 and 1931 comparing yields of different varieties of barley. In order to allow for variation in such factors as weather and soil fertility, the experiment was carried out over two years and six locations, and Fisher reported a varieties $\times$ locations $\times$ years table of mean yields. Yates and Cochran (1938) subsequently analyzed this example and noted the problems posed for generalizability by the presence of variety $\times$ location interactions.

Cochran (1983, p.69) cites another example of a series of experiments in England on the responses of sugar beets to fertilizers, conducted at 12 stations:

> After three years an argument arose for stopping the experiments because effects, while profitable, had been rather modest from year to year; the average responses to 90 lbs. nitrogen per acre were 78, 336, and 302 lbs. sugar per acre in the three years. There seemed to be little more to learn. A decision to continue the experiments was made, however, because all three years had unusually dry summers. In the next two years, both wet years, the average effects of nitrogen rose to 862 and 582 lbs. sugar per acre.

Clearly, the very meaning of a target population for generalization needs to be clear and may involve generalizations over time and time-dependent conditions.

The move from an experimental population to a target population typically involves substantial resources not possessed by individual experimenters. Cochran (1983) points to cooperative experiments on the treatment of leprosy conducted simultaneously with the same plan, treatment, and measurement of response in Japan, the Philippines, and South Africa. Such cooperative efforts are difficult to arrange. More often, following the results of an initial experiment with "significant results," other investigators carry out variants of it with different techniques, slightly different treatment variables, and different subjects in different locations (e.g. see Rosenthal and Rubin 1979), leaving open the question of generalizability of the original findings.

Many large-scale social experiments in the U.S. have used strategic variation in experimental materials to establish external validity, though they rarely use that term. For example, the negative-income-tax experiments took place in various locales that differed on such variables as urban/rural, racial composition, and female-headed families (see the discussion in Fienberg, Singer, and Tanur, 1985). To us, the ideal solution to the problem of external validity would be to sample the subjects upon which the experiment is to be performed from the populations to which the experimenter would like to generalize. Thus, the negative-income-tax experiments might have sampled poor people across the nation, the housing-allowance study might have sampled participants or cities, etc. In this way an experiment would have been totally embedded within a sampling design. The only large-scale experiment that we are aware of that was designed using a nation-wide probability sample was the Social Security disability experiment—and that was never fielded. Some disagree with this ideal sampling prescription, noting that there may not exist a meaningful population of inference, and continue to argue for strategic variation.

There is currently a movement in cognitive psychology attempting to generalize laboratory findings to larger populations through the use of large-scale surveys [e.g., see Fienberg, Loftus, and Tanur (1985) and Jabine, Straf, Tanur, and Tourangeau (1984)]. For example, in an academic laboratory, using students as subjects, Loftus and Fathi (1985) examined the order in which students recall autobiographical events that happen repeatedly. They found that when retrieving information about academic examinations, students' memories were better if they retrieved beginning with the most recent incident. This method of backward search may succeed because the first few items searched for are easier to retrieve, and thus provide a better starting point for retrieval of the entire chain. Interestingly, when retrieving health-care visits, students seemed to find it easier to recall in the forward direction (Fathi, Schooler, and Loftus 1984). This apparent discrepancy raises questions about whether retrieval strategies are specific to classes of recall tasks. In retrieving academic-examination information, for example, since examinations are fairly independent events, people might well be expected to begin by retrieving the most recent and available instance. With health-care visits, on the other hand, there is more likely to have been some causal relationship between the various visits (e.g., you broke your ankle, so you went to the orthopedic specialists, who told you to go to the radiologist for X-rays).

Important questions that remain include whether these findings on order of recall will generalize to a broader population than that of college students and to settings other than the cognitive laboratory. If they do, we shall have firmer theoretical knowledge about the working of memory as well as helpful guidance for the construction of survey instruments to be used to solicit recall of autobiographical events. An attempt to answer these questions is now underway in the form of a methodological experiment in the National Health Interview Survey/National Medical Expenditure Survey Linkage Field Test (White and Mathiowetz 1985).

The laboratory result described above is rather subtle—the more effective method of recall may be only slightly better than the less effective, and the appropriate recall strategy may vary with the type of material being recalled. But even small gains in effectiveness of recall may offer large payoffs in increased accuracy when we are dealing with large national samples and many thousands of potentially recallable events. It is in these cases of effects that are subtle and small on an individual basis

(though perhaps large in the aggregate), rather than in the cases of "slam-bang effects" (Gilbert, Light, and Mosteller 1975) whose generalizability is practically beyond question, that extensions to larger and more varied populations is crucial.

## 6. INFERENCE FOR EXPERIMENTS EMBEDDED IN SURVEYS

Issues regarding inferences from experiments embedded in surveys are rarely discussed in published sources. Depending on the perspective that one adopts, inferences can proceed in at least three different ways:

(a) By using the standard inference paradigm based on experimental randomization, which relies largely on internal validity and the assumption of unit-treatment additivity.

(b) By using the standard sampling paradigm, which, for a two-treatment experiment embedded in a survey, conceptualizes two populations of values, a pair for each unit or individual. Then inferences can be focused on the mean difference (or the difference in the means of the two populations).

(c) By conceptualizing a population of experiments of which the present embedded experiment is a unit or a sample of units.

These approaches focus on the same experimentally observed quantities but deal with the inference question differently. We illustrate them using as our example a variant on the split-ballot approach for examining differences between alternative questionnaire structures in a sample survey.

Tourangeau and Rasinski (1986) carried out an experiment to study context effects in attitude surveys, similar to the one described in Section 3.2 above. For this experiment they used 4 issues at differing levels of familiarity (abortion, welfare, aspects of banking legislation, and proposed immigration legislation) with 4 different orders of presentation of the target issues (structured using a Latin square), 2 versions of the context questions used in advance of the target question (positive or negative), and 2 methods of structuring the context questions (*mixed* across issues or *organized* by issue with context questions followed by the linked target question). This yielded 16 versions of the questionnaire, to which the investigators added 2 additional versions with neutral context questions, for a total of 18 versions. The responses of interest consisted of answers (favor/oppose or agree/disagree) to the four target issues (plus possible "don't know" responses).

Each interviewer used (approximately) a SRS of respondents from telephone banks listed in the Chicago directory. The interviewers received the questionnaires in batches of 18 and worked their way through a batch as they reached respondents willing to be interviewed (there was a 35% combined rate of refusal and nonresponse). There were 4 interviewers each of whom carried out 5 batches of 18 interviews. Thus there were a total of 360 responses. Here we ignore the nonresponse problems and treat the sample as if it consisted of all selected respondents.

We can consider the 4 interviewers is blocks and within each block we have 5 replications of an 18-treatment experiment, where 16 of the treatments represent a $4 \times 2 \times 2$ factorial design. The outcomes for a given interview $\times$ treatment combination can be cross-classified according to the 4 dichotomous target response variables. Because of this categorical response structure, Tourangeau and Rasinski analyzed the "effects" measurable by this overall design using logit models.

How do the three approaches to inference differ for this experiment? Method (a)

treats the outcomes in the traditional experimental fashion, with the block effects due to interviewer taken as fixed, and using up 3 d.f. (but see below). The 18 treatment combinations would be used to estimate various main effects and interaction effects involving context (although the power to detect interactions may not be very substantial). The block × treatment interaction would typically go into the "error term" in such an analysis, although specific components of the interaction could be examined in the multivariate logit model. This approach makes inferences internal to the experiment, although the study was clearly designed to generalize to the broader implications of such context effects. This analysis is based on a likelihood approach to modelling, in contrast to an approach to inference solely via the randomization features of the design.

Method (b) treats every respondent in the population as having a "potential" response to each experimental condition, attempts to estimate the population proportions of respondents falling into the $2^4$ response categories for each treatment combination, and then compares those estimated proportions in order to measure various "effects". From this perspective, we are using the sample survey as if it consisted of 18 different SRSs, and we are not so much interested in the internal structure of the experiment as we are in how the separate internal parts "represent" the corresponding populations. This approach stumbles over interviewer effects, since it ignores them.

Method (c) can be viewed, in part, as a way out of the dilemma that interviewers pose for the sampling approach in method (b). Here we treat the experiment actually done in Chicago as if it were a sample (of size 1) from a universe of possible experiments, and here the interviewers are thought of as a sample from a population of interviewers. Thus we could, from a model-based perspective, think of the interviewers as leading to a random effect in the analogue of a mixed-model analysis of variance, and then we would use the interviewer × treatment interaction term as the relevant error component.

We have felt it important to illustrate the three modes of inference with a concrete example—but such concreteness has its price. In particular, some might object to the analysis illustrating approach (a) and using interviewers as a fixed factor. The distinction between fixed and random factors is at best a fuzzy one. Indeed, a classic example given in Scheffé (1959, p.261) uses machines as fixed because the experimenter is interested in the individual performance of the machines, while workers are random, regarded as a random sample from a large population. It would seem easy enough to reverse that thinking and consider workers fixed because they constitute a permanent work force and machines as random because they are a sample from a population of machines that might be purchased as replacements. Nonetheless, the experimental randomization only provides a formal justification for internal inferences and thus for a fixed-effects analysis, and it is this structure that approach (a) is considering. Any additional randomness is in the eye of the analyst, and constitutes an issue of generalization (in the sense of external validity) and not internal analysis to establish internal validity. Considering certain effects as random is reminiscent of the notion of sampling levels of treatment discussed in Section 2. One could justify the stance taken in (a) by the fact that the interviewers taking part in the experiment would continue as part of the NORC work force: if we consider them randomly sampled from that work force or from some larger population, then we can more easily assume a random-effects model. Moreover, method (b) would, if interviewers were indeed sampled, come much closer to method (c). Random-effects

models have not received much direct attention in the sampling literature (see Fienberg and Tanur 1987).

There is no single "correct" way to view inference for experiments embedded in surveys, and the purpose of this discussion is to initiate a more careful look at the different perspectives one might consider adopting on the inference question. Illustrative empirical analyses would shed light on any differences in substantive conclusions stemming from the different perspectives.

## 7. CONCLUSION

As we have explored the many examples of intertwining of experimentation and sampling detailed here, we have been amazed at the amount of lamination we have been able to point out in the design stage. Experiments embedded in sample surveys use sampling to choose treatment combinations. Sampling to measure outcomes is embedded in experiments that are embedded in a higher-order sampling structure for the sake of generalization. We have been surprised in a different way, however, as we examined the analyses proposed or carried out in these hybrid studies. All too often we note features carefully embedded in the design stage are not fully capitalized upon in analysis. The separation of the statistical subspecialties dealing with experimentation and sampling exacts a heavy toll from the practitioners of both. The use of analyses that are less powerful than they could be for experiments embedded in surveys is part of that toll.

## REFERENCES

Aronson, E.; Brewer, M., and Carlsmith, J.M. (1985). Experimentation in social psychology. *Handbook of Social Psychology. Volume 1* (G. Lindzey and E. Aronson, *eds.*). Third Edition. Random House, New York.

Bailar, B.A. (1983). Interpenetrating subsamples. *Encyclopedia of Statistical Sciences* (S. Kotz and N. Johnson, *eds.*), Wiley, New York, Volume 4, 197–201.

Bailar, B., and Biemer, P. (1984). Some methods for evaluating nonsampling error in household censuses and surveys. *W.G. Cochran's Impact on Statistics* (P.S.R.S. Rao and J. Sedransk, *eds.*). Wiley, New York, 253–275.

Bailey, L.; Moore, T.F., and Bailar, B.A. (1978). An interviewer variance study for the eight impact cities of the National Crime Survey cities sample. *J. Amer. Statist. Assoc.*, 73, 16–23.

Campbell, D.P. (1957). Factors relevant to the validity of experiments in social settings. *Psychol. Bull.*, 54, 297–312.

Campbell, D.P., and Stanley, J.C. (1963). Experimental and quasi-experimental designs for research. *Handbook of Research on Teaching* (N.L. Gage, *ed.*), Rand McNally, Chicago, 171–246.

Clapham, A.R. (1931). Studies in sampling technique: Cereal experiments. I. Field technique. *J. Agricultural Sci.*, 21, 366–371.

Cochran, W.G. (1968). Errors of measurement in statistics. *Technometrics*, 10, 637–666.

Cochran, W.G. (1983). *Planning and Analysis of Observational Studies* (L.E. Moses and F. Mosteller, *eds.*), Wiley, New York.

Cochran, W.G., and Cox, G.M. (1957). *Experimental Designs*. Second Edition. Wiley, New York.

Cochran, W.G., and Watson, D.J. (1936). An experiment on observer's bias in the selection of shoot heights. *Emperical J. Experimental Agriculture*, 4, 69–76.

Cook, T.D., and Campbell, D.P. (1979). *Quasi-Experiments: Design and Analysis Issues for Field Settings.* Rand McNally, Skokie, Illinois.

Durbin, J., and Stuart, A. (1951). Differences in response rates of experienced and inexperienced interviewers. *J. Roy. Statist. Soc. Ser. A*, 114, 163–206.

Fathi, D.C.; Schooler, J., and Loftus, E.F. (1984). Moving survey problems into the cognitive psychology laboratory. *Proceedings of the American Statistical Association Section on Survey Research Methods*, Amer. Statist. Assoc., Washington, 19–21.

Federer, W.T. (1976a). Sampling, blocking, and model considerations for the completely randomized, randomized complete block, and incomplete block experimental design. *Biometrical J.*. 18, 511–525.

Federer, W.T. (1976b). Sampling, blocking, and model considerations for the $r$-row and $c$-column experimental designs. *Biometrical J.*, 18, 595–607.

Federer, W.T. (1977). Sampling, blocking, and model considerations for split plot and split block designs. *Biometrical J.*, 19, 181–200.

Fellegi, I.P. (1964). Response variance and its estimation. *J. Amer. Statist. Assoc.*, 59, 1016–1041.

Fellegi, I.P. (1974). An improved method of estimating the correlated response variance. *J. Amer. Statist. Assoc.*, 69, 496–501.

Fienberg, S.E.; Loftus, E.F., and Tanur, J.M. (1985). Cognitive aspects of health survey methodology: An overview. *Milbank Memorial Fund Quart.*, 63, 547–564.

Fienberg, S.E.; Singer, B., and Tanur, J.M. (1985). Large scale social experimentation in the U.S.A. *A Celebration of Statistics: The ISI Centenary Volume* (A.C. Atkinson and S.E. Fienberg, *eds.*), Springer-Verlag, New York, 287–326.

Fienberg, S.E., and Tanur, J.M. (1985). A long and honorable tradition: Intertwining concepts and constructs in experimental design and sample surveys. *Bull. Internat. Statist. Inst.*, Book II, 10.1-1–10.1-18.

Fienberg, S.E. and Tanur, J.M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *Internat. Statist. Rev.*, 55, 75–96.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh.

Fisher, R.A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.

Gilbert, J.P.; Light, R.J., and Mosteller, F. (1975). Assessing social innovations: An empirical base for policy. *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs* (C.A. Bennett and A.A. Lumsdaine, *eds.*), Academic Press, New York, 39–193.

Groves, R.M., and Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quart.*, 50, No. 2, 251–266.

Hansen, M.H.; Hurwitz, W.N., and Bershad, M.A. (1961). Measurement errors in censuses and surveys. *Bull. Internat. Statist. Inst.*, 38, 359–374.

Hansen, M.H.; Hurwitz, W.N., and Madow, W.G. (1953a). *Sample Survey Methods and Theory. Volume I.* Wiley, New York.

Hansen, M.H.; Hurwitz, W.N., and Madow, W.G. (1953b). *Sample Survey Methods and Theory. Volume II.* Wiley, New York.

Hansen, M.H.; Hurwitz, W.N.; Marks, E.S., and Mauldin, W.P. (1951). Response errors in surveys. *J. Amer. Statist. Assoc.*, 46, 147–190.

Jabine, T.B. and Rothwell, N.D. (1970). Split-panel tests of census and survey questionnaires. *Proceedings of the American Statistical Association Social Statistics Section*, Amer. Statist. Assoc., Washington, 4–13.

Jabine, T.B.; Straf, M.; Tanur, J.M., and Torangeau, R. (*eds.*) (1984). *Cognitive Aspects of Survey Methodology: Building a Bridge between Disciplines*. National Academy Press, Washington.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *J. Amer. Statist. Assoc.*, 57, 92–115.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples. J. Roy. Statist. Soc. Ser.B, 36, 1-37.

Loftus, E.F., and Fathi, D. (1985). Retrieving multiple autobiographical memories. *Social Cognition*, 3, 280–295.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *J. Roy. Statist. Soc.*, 109, 325–378.

Mosteller, F. (1978). Nonsampling errors. *International Encyclopedia of Statistics. Volume I* (W.H. Kruskal and J.M. Tanur, *eds.*), Free Press, New York, 208–229.

Panel on Privacy and Confidentiality as Factors in Survey Response, Committee on National Statistics (1979). *Privacy and Confidentiality as Factors in Survey Research*. Nat. Acad. Sci., Washington.

Robinson, D.L. (1987). Estimation and use of variance components. *The Statistician*, 36, 3–14.

Rosenthal, R., and Rubin, D.B. (1979). Issues in summarizing the first 345 studies of interpersonal expectancy effects. *Behavioral and Brain Sci.*, 3, 410–415.

Rossi, P.H., and Anderson, A.B. (1982). The factorial survey approach: An introduction. *Measuring Social Judgments* (P.H. Rossi and S.L. Nock, *eds.*), Sage, Beverly Hills, California, 15–67.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.

Schuman, H.; Steeh, C., and Bobo, L. (1985). *Racial Attitudes in America: Trends and Interpretations*. Harvard Univ. Press, Cambridge, Massachusetts.

Scott, C. (1961). Research on mail surveys. *J. Roy. Statist. Soc. Ser.* A, 124, 143–205.
Scott, C. (1973). Experiments on recall error in African household budget surveys. Unpublished paper presented at meeting of International Association of Survey Statisticians, Vienna, Austria (August 1973).
Stokes, S.L. (1986). Estimation of interviewer effects in complex surveys with application to random digit dialling. *Proceedings of Second Annual Research Conference*, U.S. Bureau of the Census, Washington, 21–31.
Tourangeau, R. (1986). Personal communication.
Tourangeau, R., and Rasinski, K.A. (1986). Context effects in attitude surveys. Unpublished manuscript.
Waksberg, J., and Pearl, R.B. (1965). New methodological research on labor force measurement. *Proceedings of the Social Statistics Section*, Amer. Statist. Assoc., Washington, 227–237.
White, A.A., and Mathiowetz, N. (1985). The National Health Interview Survey as a sampling frame for a National Medical Expenditure Survey: A field test design. Paper presented at the annual meeting of the American Statistical Association, Las Vegas.
Wilk, M.B., and Kempthorne, O. (1956). Some aspects of the analysis of factorial experiments in a completely randomized design. *Ann. Math. Statist.*, 27, 950–985.
Woltman, H.F.; Turner, A.G., and Bushery, J.M. (1980). comparison of three mixed-mode interviewing procedures in the National Crime Survey. *J. Amer. Statist. Assoc.*, 75, 534–543.
Wood (A.J.) Research Corporation (1959). *Woodchips*, 4, No. 1.
Yates, F. (1935). Some examples of biased sampling. *Ann. Eugenics*, 6, 202–213.
Yates, F. (1981). *Sampling Methods for Censuses and Surveys*. Fourth Edition. Macmillan, New York.
Yates, F. (1985). Book review of "W.G. Cochran's Impact on Statistics" (P.S.R.S. Rao and J. Sedransk, *eds.*), *Biometrics*, 41, 591–592.
Yates, F., and Cochran, W.G. (1938). The analysis of groups of experiments. *J. Agricultural Sci.*, 28, 556–580.
Yates, F., and Zacopany, I. (1935). The estimation of the efficiency of sampling with special reference to sampling for yields in cereal experiments. *J. Agricultural Sci.*, 25, 545–577.

*Received 12 January 1987*
*Revised 7 August 1987*
*Accepted 12 January 1988*

*Department of Statistics*
*Carnegie Mellon University*
*Pittsburgh, PA 15213*
*U.S.A.*

*Department of Sociology*
*State University of New York at Stony Brook*
*Stony Brook, NY 11794*
*U.S.A.*