International Statistical Review (1987), **55**, 1, pp. 75–96. Printed in Great Britain

Experimental and Sampling Structures: Parallels Diverging and Meeting

Stephen E. Fienberg[1] **and Judith M. Tanur**[2]

[1]*Departments of Statistics and Social Sciences, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 15213, USA.* [2]*Department of Sociology, State University of New York, Stony Brook, New York 11794, USA*

Summary

The design and analysis of *randomized* experiments and *randomly selected* sample surveys are traced to the work of Fisher, Neyman and Tchuprov in the 1920's and 1930's, although precursors to their work appeared many years earlier. This paper explores some of the developments flowing from their pioneering efforts with an emphasis on the parallels between the methodologies. After reviewing the basic parallels between concepts in the design of experiments and the design of sample surveys, the paper turns to a new class of parallels linking restricted forms of sampling to the design-of-experiments literature on treatment structures, such as that on balanced incomplete block designs. The parallel concepts for control in the two areas lead naturally to a discussion of embedding experiments in surveys or surveys in experiments. After speculating on the possible causes of the separation of the areas, the paper summarizes the parallel controversies between the two modes of inference, design-based and model-based, used in both the experimental design and sample survey literatures. In summary, the paper proposes how new intertwining concepts and constructs may emerge in future research and enrich future practice.

Key words: Control; Experimental design; Embedding; Randomization; Restricted randomization; Sampling design.

1 The tradition begins

The design of *randomized* experiments and the use of *random selection* in sampling are usually traced to the work of Fisher and Neyman in the 1920's and 1930's (see the related discussions in the biographies by Box (1978) and by Reid (1982) respectively) as well as to Tchuprov (1923), although precursors to their work appeared many years earlier; see, for example, the discussion by Seng (1951) and Zarkovich (1956, 1962). In the earlier work, randomization and random selection were primarily associated with the notions of fairness, objectivity, and, even later, representativeness (Fienberg, 1971; Kruskal & Mosteller, 1980). Smith & Sugden (1985) review the pivotal role the International Statistical Institute played in some of these early discussions in the area of sampling. The novel departure in the work of Fisher, Neyman and Tchuprov was the introduction of chance mechanisms in order to make available probability-based methods of inference at the analysis stage. In the present paper we trace some of the developments flowing from this early work, noting in particular that several statisticians (e.g. Cochran, Finney, Hartley, Madow, Yates) contributed to the literature in both areas, and so it is not surprising that there are commonalities across the methodologies.

While we argue that the two traditions grew up together, Stephan (1948), writing almost 40 years closer to the events, pointed out a lack of communication at the outset very similar to that which we point out here. In a presentation before the 25th Session of the International Statistical Institute, he argued (Stephan, 1948, p. 30)

> ... developments in agriculture and engineering have both direct and indirect effects on sampling survey practice. They provided principles of design and contributed to the growth of applied mathematical statistics. Still there were many practical problems and obstacles that delayed the immediate extension of the methods developed for field trials and manufacturing to large-scale surveys. One of the obstacles was the relative lack of communication between statisticians engaged in different types of work.

Stephan went on to suggest that it is institutional mechanisms that overcome communication barriers and encourage cross-fertilization. It is to further encourage cross-fertilization that we have embarked on the present research.

In § 2, we briefly review the basic parallels between the design of randomized experiments and sampling studies. We include a detailed description of a two-treatment randomization design for an experiment and show that the structure is identical to the one that describes the selection of a simple random sample. In § 3, we go on to note some of the more modern parallels that have been developed. One of the more important lessons to be learned from the intertwining concepts and constructs of experimentation and sampling is that the two can profitably be combined, with sampling embedded in experiments and experiments embedded in sampling structures. In § 4, we pursue this theme, reviewing the institutionalization of the embedding of experiments within samples, including Mahalanobis' concept of interpenetrating networks of samples and voluminous work at the U.S. Bureau of the Census, especially in connection with the evaluation of decennial census methodology. In contrast to some of this careful work, we point to examples in which investigators have failed to take full advantage of the possibilities of control offered by the device of embedding.

Why is it that modern researchers and students seem to be ignorant of these parallels across fields? In § 5, we propose some tentative answers to this question. Then in § 6, we summarize the two modes of inference that are used in both the experimental design and sampling literatures: design-based inference, which relies on the probabilistic structure associated with the design; model-based inference, which introduces stochastic components as part of parametric structures. We take special note of the parallel controversies in the two areas. In the concluding section, we speculate on how new intertwining concepts and constructs may emerge in future research and may enrich future practice.

2 Basic parallels

It is well known that the basic concepts in the design of sampling studies parallel those for the design of randomized experiments. For example, coupled to the notion of randomization in experimentation is probability (random) sampling, both involving the introduction of chance mechanisms (for assignment of treatments to units in experiments and for the choice of sample units in surveys) in order to make available probability-based methods of inference at the analysis stage. The parallel concepts and structures are most easily illustrated in the simple two-treatment or two-group experiment and its parallel structure, the simple random sample.

Consider a universe of N objects, $\mathbf{U} = \{u_1, u_2, \ldots, u_N\}$, and a sample selection function $\mathbf{A}_s = (A_1, A_2, \ldots, A_N)$, where

$$A_i = \begin{cases} 1 & \text{if } i \in T_1, \\ 0 & \text{if } i \in T_2. \end{cases}$$

In a two-treatment experiment, the sample selection function, $\mathbf{A}_s$, specifies which members of the universe are allocated to treatment 1, that is T_1, and which to treatment 2, that is T_2. In the sampling situation, allocation to T_1 corresponds to being selected for inclusion in the sample, and allocation to T_2 corresponds to nonselection. If T_1 contains n members, then experimental randomization and simple random sampling both take each of the

$$\binom{N}{n}$$

$\mathbf{A}_s$'s with n of the A_i equal to 1 to have probability of selection equal to

$$1\Big/\binom{N}{n}.$$

In an experiment, under the null hypothesis of no differential treatment effect, the observed value of the test statistic (e.g. the difference in sample means) is compared with the distribution of all

$$\binom{N}{n}$$

possible values associated with the totality of allocations that could have been obtained under the randomization. This use of what is now known as randomization theory originated in the work of Fisher (1925, 1926), and it figures prominently in his 1935 book, *The Design of Experiments.* Fisher's theory, as it was later developed by Kempthorne (1952, 1955) and others, utilizes the formal act of randomization in exactly the same way that the standard approach to survey analysis, originally proposed by Tchuprov (1923) and Neyman (1934) and developed further by Hansen, Hurwitz & Madow (1953a,b), utilizes random selection in sampling.

We note that, while the language is the same, some of the purposes of the randomization structures in the sampling and the experimental contexts *are* different. For example, in the simplest experiment, we are trying to compare the effects of two treatments. In a sampling study, on the other hand, we want to generalize from one group to the other, i.e. from the sample to the rest of the population. (The sampling literature usually speaks of generalizing to the entire population rather than to the rest of the population, i.e., minus the sample.) As Bartlett (1978) points out, Fisher stressed that in controlled experiments there is the opportunity for deliberately introducing randomness into the design in order to separate systematic variation from purely random error. In an experiment, through randomization we 'hold everything constant', and thus we can attribute any effects to the treatment differences; in the sampling context, the random selection and the fact that no treatment is applied to the sampled group allows us to make the generalization to the rest of the population. Nonetheless in both contexts, the randomization structure is used to provide a meaningful estimate of variability. In the experimental context, this underlying variability is the yardstick by which we compare the measurements of the responses to the treatments; in the sampling context, the sampling variability induced by the randomization is used to gauge the precision of sample estimates of population quantities.

The use of homogeneous groups is common to both experimental design and to sampling design. Homogeneous groups are used in experimental design to minimize experimental error via the device of blocking (Cochran & Cox, 1957, p. 106ff.), each replication being carried out on a homogeneous group of subjects. Unlike randomization which attempts to control for other factors by ensuring that each treatment has an equal chance of being favoured or handicapped by an extraneous source of variation, blocking exerts its control by attempting to segregate the effects of an extraneous source of variation and thereby reduce experimental error. Similarly, homogeneous groups are used in sampling to minimize sampling error via the device of stratification (Cochran, 1977, p. 89ff.), with samples drawn from each of the homogeneous groups into which a population is divided. Note that this analogy is particularly strong in design, where the blocking and the stratification are both used as control structures, but is less strong in analysis where the error terms for the two techniques differ. In randomized blocks, the error term is defined as the block-by-treatment interaction, while in stratification there is only one treatment (we examine only those in the sample) and thus the error term is 'within replications'. Therefore, the real analogy is between stratification and randomized blocks with multiple replications within blocks.

Devices in experimental design that aim to reduce experimental error by simultaneously controlling for two or more sources of extraneous variability, such as Latin and Graeco–Latin squares (Fisher, 1935, Ch. V; Cochran & Cox, 1957, p. 117ff.), find parallels in sampling design. Just as these procedures are used in experimental design when the pairing of all possible combinations of control factors is impossible, when there are two or more dimensions of stratification and choosing a sample from each cell in the cross-classification is unwieldy, the application of Latin or Graeco–Latin squares produces a method for choosing strata to include in the sample and is called lattice sampling (Cochran, 1977, p. 228ff.) or 'deep' stratification (Frankel & Stock, 1942; Tepping, Hurwitz & Deming, 1943; Hansen et al., 1953a, p. 480 ff.; Kish, 1965, pp. 488–495).

At a more refined level, the convenience of split-plot designs (Cochran & Cox, 1957, p. 293ff.) is echoed in the analogous sampling technique, cluster sampling (Hansen et al., 1953b, Ch. 6; Cochran, 1977, p. 233ff.). In a split-plot experiment we can think in terms of two sources of error: one between plots and one within plots. Similarly, in cluster sampling, we can think in terms of two components of variability, one between clusters and one within clusters. In the experimental context, separating out the between-plot component of variability allows for greater precision in sub-plot comparisons, whereas in cluster sampling, because the sample is used to produce estimates of overall population quantities, the two components are combined to produce an overall sampling variance which is larger than that associated with a simple random sample of the same size. In the analysis phase, covariance analysis (Cochran & Cox, 1957, p. 82ff.) in an experimental investigation adjusts estimates of the magnitude of treatment effects for environmental influences in the same way that post-stratification and regression estimates (Cochran, 1977, p. 189ff.) are used to adjust sampling results.

The analysis of variance (ANOVA) structure is used in both areas as a way of summarizing information associated with many of the basic methods for control, although this usage is found in the sampling literature primarily in the work of authors steeped in the traditions of both areas; Yates (1985) attributes this usage to Fisher. This use of the analysis of variance is often related to a Model I or fixed effects linear model with normally distributed error term, although such a link is not the only possible formalization for inference purposes. There are, in addition, analogues for the experimental-design-based Model II or random effects linear models in the sampling context. Model II approaches have not received much attention in the sampling literature, primarily because of the heterogeneity among the units of the typical sampling population. We note two exceptions. The conceptualization of models for the total survey error can take the component of variation due to interviewer as a random effect (Hartley & Rao, 1978;

Table 1

Parallels between basic concepts in design and analysis of experiments and in sampling design and analysis

Experiments	Sampling
Randomization	Random sampling
Blocking	Stratification
Latin squares	Lattice sampling (deep stratification)
Split-plot designs	Cluster sampling
Covariance adjustment	Post-stratification

Hansen et al., 1951). In small area estimation, components of variance approaches seem appropriate because the assumption that homogeneity holds within small areas is less problematic than that it holds across large and disparate areas; for example, see the various papers in the book by Platek et al. (1986).

In order to confirm the existence of these parallels and to suggest others, we reviewed many of the basic textbooks in experimental design and sampling to see whether the parallel structures were referenced or used as pedagogic tools. The textbooks on experimental design exhibited virtually no direct reference to this parallel structure (a notable exception being a passing reference by Cox (1958)) although the reader perusing Kempthorne's (1952) book will find formulae of direct use in a sampling context and even a discussion of sampling within experiments (see § 4 below). When we looked at sampling texts, we found a parallel neglect, but with some more exceptions. Cochran (1977) uses the analysis of variance structure throughout as a summary device, while Hansen et al. (1953a) and Kish (1965) discuss the parallel between Latin squares and lattice sampling. A more fundamental exception is the text by Yates (1981), which is replete with cross-referencing between the areas.

To summarize, there are several basic concepts in the design and analysis of experiments which have exact parallels in sampling design and analysis. They include those in Table 1. This list is far from definitive. Similar parallels can be found in work on allocation and optimal design in the experimental and survey literature. We note that the absence of treatments in the sampling context means that there is no immediate role there for analogues of the factorial treatment structures that dominate much of the experimental literature. There are, however, some less-than-immediate parallels, as we note in § 3.

3 More modern parallels: Restricted randomization

Blocking and Latin squares were introduced in agricultural field experimentation in order to control for known heterogeneity in the plots. But for some layouts within these classes of designs a high proportion of contiguous plots can inadvertently receive identical treatments (for example, down diagonals in Latin squares). Forms of *restricted randomization* have been proposed to avoid this problem; see, for example, Yates (1948), Grundy & Healy (1950), Holschuh (1980) and Youden (1964, 1972). Through the consideration of a particular example, Bailey (1985) has demonstrated that restricted randomization can serve the role of both blocking and ordinary randomization (as described above) while at the same time controlling for spatial arrangements of plots. This notion of restricted randomization also has applicability in the sampling context, e.g. to control for the geographical spread of a sample or more generally to eliminate the possibility of 'bad' samples. Because the typical human population of interest in sampling is large and heterogeneous, the simple devices for restricted randomization in experimentation cannot be carried over directly.

In this section we focus on an innovative approach to the use of restricted randomization in a sampling context that relies on a different parallel with experimental design ideas. This approach originated in work by Chakrabarti (1963), and it relates the notion of the *support of a sampling plan* to the structure of *balanced incomplete block designs* (BIBD). The *support of a sampling plan* consists of those samples having positive probability of selection. For example, in simple random sampling (SRS), if N is the population size and n the sample size, the support consists of all

$$\binom{N}{n}$$

possible samples. The research objective in this work is to develop sampling plans equivalent to SRS with respect to 1st-order and 2nd-order inclusion probabilities for each unit and pair of units from the population, but which have smaller support than SRS. If the sampling plan has 1st-order inclusion probabilities equivalent to those of SRS, that is $\pi_i = n/N$, then it will produce samples with means that are unbiased estimates of the population mean. If, in addition, the sampling plan has 2nd-order inclusion probabilities equivalent to those of SRS, that is $\pi_{ij} = n(n-1)/N(N-1)$, then the sample means will have variances equal to those of sample means produced by SRS. The motivation for this objective comes not only from the idea of restricted randomization that eliminates possible samples that are not 'representative' without sacrificing the gains achieved by randomization, but also from the desire to reduce field expenses associated with data collection through an implicit trade-off between stratification and clustering.

We recall the definition of a BIBD with $bk = rt$ experimental units:

(i) each of b blocks contains k units;
(ii) each of t treatments occurs the same number of times, r, in all;
(iii) every pair of treatments occur together, in the same block, the same number of times, λ.

Once we identify the similarity of pairs of treatments in a block and pairs of units in a sample the new parallel structure in Table 2 falls into place.

The parallel structure here links samples with the block structure rather than with the treatment structure as in the case in our basic list of parallels in § 2. The key result based on this new parallel was given by Chakrabarti (1963) as follows.

THEOREM 1. *A sampling plan with uniform selection probabilities over the samples in the support is equivalent to* SRS *with respect to all* 1*st- and* 2*nd-order inclusion probabilities if and only if it is associated with a* BIBD *with* $N = t$ *and* $n = k$, *which has distinct blocks.*

One of the implications of this result is that, in order to develop complex sampling plans with smaller support than SRS, we constantly need to keep in mind the BIBD structure, even if we cannot write out the actual design for very large N and substantial sample size, n.

Table 2

Parallels between pairs of treatments in a block and pairs of units in a sample

BIBD	Sampling plan
Blocks	Samples
Treatments	Units
Block size, k	n, sample size
# treatments, t	N, # population units
# blocks, b	Support size for sampling plan

Wynn (1977) then showed Theorem 2.

THEOREM 2. *For any sampling plan P_1, consisting of samples of size n from a universe of size N, there is another plan P_2 equivalent to P_1 with respect to all 1st- and 2nd-order inclusion probabilities with support size no greater than $N(N-1)/2$.*

This result, however, is not constructive, i.e. it does not show us how to achieve the reduction. But it can be used along with the following theorem, also due to Wynn (1977), to set bounds on the improvement achievable using results like Theorem 1.

THEOREM 3. *There is a sampling plan with support size N, equivalent to* SRS *with respect to 1st- and 2nd-order inclusion probabilities, if and only if there exists a symmetric* BIBD *with $t = b = N$. No such plan exists with support size less than N.*

This result implies that the number of possible samples in the support of the sampling plan can be reduced from

$$\binom{N}{n}$$

to between

$$\binom{N}{2}$$

and N by careful structuring and balancing. Since N is typically very large we thus retain a potentially sound basis for large-sample approximations to design-based inference methodology. Wynn also considers convexity properties of sets of sampling plans, and notes the analogy of the resulting upper bounds on the minimum number of samples with results in theory of optimal design (Kiefer, 1961).

Foody & Hedayat (1977) carry this parallel one step further, looking at BIBD's with repeated blocks. (This removes the restriction in Chakrabarti's theorem regarding distinct blocks). By allowing for repeated blocks, one can select specific blocks a maximal number of times and thereby maximize, in the sampling context, the probability of selection of certain samples while still preserving the 1st- and 2nd-order inclusion probabilities. The practical value of such sampling plans is especially apparent when government statistical agencies designing elaborate surveys need to ensure the presence of sample units in selected political jurisdictions that are the constituencies of elected representatives who must approve the budgets for the surveys. In the United States, such considerations have been formally built into the decennial redesign of the Current Population Survey.

If we let b^* be the number of distinct blocks (i.e. the support size), then for SRS the corresponding BIBD has

$$\lambda = \binom{t-2}{k-2}, \quad b^* = b = \binom{t}{k}, \quad r = \binom{t-1}{k-1}.$$

When not all blocks are distinct $b^* < b$. Foody & Hedeyat (1977) provide a necessary and sufficient condition for the existence of BIBD's with the above parameters but where $b^* < b$. An example of one of the consequences of their result is Theorem 4.

THEOREM 4. *Suppose there exist two* BIBD's *with the same set of t treatments, the same block size, k, and with nonoverlapping supports of sizes b_1^* and b_2^*, respectively. Then there exists a* BIBD *with $t' > t$ treatments, support size*

$$b^* = \binom{t'}{k} - b_1^*,$$

and total number of blocks,

$$b = e\binom{t'}{k},$$

where $e = \lambda_2/gcd(\lambda_1, \lambda_2)$ *and* λ_1 *and* λ_2 *are the number of pairs of treatments occurring together in the original* BIBD'S.

Balanced incomplete block designs with repeated blocks meet the criteria of Theorem 2, and are subject to a lower bound different from that of Theorem 3 as shown by Foody & Hedayat (1977) and Hedayat (1979).

THEOREM 5. *Consider a sampling plan whose support size* b^* *is minimum among all sampling designs equivalent to* SRS *with respect to first- and second-order inclusion probabilities. Then*

$$\binom{N}{2}\Big/\binom{n}{2} \leqslant b^* \leqslant \binom{N}{2}.$$

By allowing for overlapping supports in the two original BIBD's one can still reduce the support of the BIBD for the larger number of treatments, but not as much. To date the focus in the experimental design literature on the consequences of these and related results has been on problems with blocks of size 3, not directly applicable to large-scale sample surveys. Applying these ideas to the survey area is nonetheless of interest. One way to restrict the number of samples in the support of a sampling plan is to begin with some form of stratification, and then to guide selection of units within strata according to some additional stratification variable, but not necessarily using the 'deep' stratification that parallels the Latin square control structure. Goodman & Kish (1950) suggest a mechanism for doing this called *controlled selection* which is especially useful in multistage sampling. They use controlled selection to select a small number of first-stage sampling units, and would typically use clustering at the final stage. Their approach provides for an approximate balancing with respect to the control factors; see the recent discussion of controlled sampling of Causey, Cox & Ernst (1985). Yet increased stratification changes 2nd-order inclusion probabilities and reduces the variability of estimates so that the exact relationship of controlled selection to the preceding theorems is unclear; see the discussion of Avadhani & Sukhatme (1973). There is also a link to Morris's (1979) finite selection model (FSM) for assigning treatments in an experiment by the 'balanced selection' of subjects from the finite population available for assignment, where the balancing is carried out with respect to a large number of covariates, and to Bellhouse's (1986) related approach to balancing on covariates. Bellhouse (1984b) summarizes criticisms of other model-based approaches to balancing.

Other examples of new parallels emerging in the literature are closely tied to the concepts of optimal design. For example, Meeden & Ghosh (1983) use techniques from the decision-theoretic literature on the comparison of experiments to find admissible procedures for some sampling problems. Similarly, Cheng & Li (1983, 1985) propose a class of optimality criteria drawn from optimal experimental design theory to show the link between the minimax criterion and more traditional sampling strategies suggested by Rao, Hartley & Cochran (1962). For sampling units structured according to a spatial ordering, Bellhouse (1977, 1984a) describes a one-to-one correspondence between optimal designs in an experimental context and optimal randomization schemes in a sampling setting. Finally, Iachan & Jones (1984) propose a general class of rotation sampling designs by relying heavily on ANOVA models and related ideas on the optimal search for BIBD's.

Yet additional applications of experimental design approaches in a sampling setting include the use of linked block designs for sampling on two occasions with overlapping samples (Singh & Raghavarao, 1975), the use of supplemented block and BIBD designs as an alternative approach to the randomized response method for dealing with sensitive questions in a survey context (Raghavarao & Federer, 1979), and the use of combinatorial design theory for the representation of general sampling measures (Srivastava, 1985).

4 Embedding

The parallel concepts for control lead naturally to the notion of embedding experiments in sampling studies or sampling in experiments. Traditional experimental design incorporated the idea of sampling within experimental units (for example, when interest focused on yield in an agricultural setting or on some characteristic of the members of a batch in an industrial setting) and considered sampling of experiments in order to generalize to larger populations. Early papers, for example by Yates & Zacopanay (1935) and Cochran & Watson (1936), illustrate these features of embedding in an agricultural context. The Cochran–Watson paper describes a Latin square experiment to determine the extent of observer bias (a sampling issue) in the selection of shoot-heights (studied in an experimental context)! These and other embedded instances present clues as to how the links between surveys and experiments can be made more obvious and more useful.

While sampling to measure the outcome of an experiment was an intrinsic part of the teachings of Fisher and of practice in agricultural experiments, sampling yoked with experimental design was much more rare in other fields. Speaking of the social sciences, Kish (1959, p. 333) writes

> In fact, the question of sampling, of making the experimental results representative of the specified population, have been largely ignored in experimental design until recently.

He cites the then recent work of Wilk & Kempthorne (1956) and of Cornfield & Tukey (1956), as being first steps in the introduction of sampling to experimental design. What Kish is referring to is the tendency of many experimenters in the social sciences to use groups at hand, often college sophomores, to experiment on. The implied assumption, that there is no interaction between type of subject and treatment, has been questioned frequently; for example, see Cronbach (1957) and the subsequent work in psychology demonstrating the existence and importance of such interactions. In response, some social scientists argue that, unless there is a plausible rationale leading to the theoretical expectation of such an interaction, an experiment conducted within a single population or on a few populations chosen strategically to exemplify a range of variation solves the problem. But the same issues of estimation of population quantities that dictate random selection in the sampling context should carry over to generalization of experimental results (Fienberg & Tanur, 1986b). It is in large-scale social experiments as carried out in the United States in the last several decades that these sampling considerations for subjects of experiments have become crucial; see Fienberg, Singer & Tanur (1985) for details on these experiments.

The examples given above are primarily concerned with the contributions of sampling to the structuring of experimental problems. One notable example going in the opposite direction, due to Mahalanobis (1946), is the method of *interpenetrating networks of samples* (IPNS) which provides a built-in replication structure for validating sampling results. For example, in a survey on the economic conditions of factory workers in an industrial area of India, Mahalanobis divided the area into 5 subareas. Instead of assigning each subarea to a different interviewer, he 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, Ch. 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 (Kish & Frankel, 1974).

If experiments within surveys are to be of value, one must apply the experimental principles of local control to ensure that actual differences will be detected with high probability. In large-scale sampling studies it has long been recognized that one of the largest sources of response error is associated with interviewer variability (Hansen et al., 1953a, Ch. 12), and the classic model used by the U.S. Bureau of the Census to measure the effects of interviewers on sample estimates and the resulting mean squared errors draws both explicitly and implicitly on random effects models from experimental design and ANOVA (Hansen et al., 1951). This suggests to those familiar with the experimental principle of local control that a useful way to embed an experiment within a survey would be to use interviewers as a form of block.

When one of us suggested this approach 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. Surely interviewer training and supervision must be very careful if an experimental strategy of blocking on interviewers is 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 U.S. Bureau of the Census methodological studies in the 1940's and 1950's. In one well documented study, for example, Waksberg & 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 split-panel tests with surviving documentation conducted by the Census Bureau from 1957 through 1969 the only one, according to the description of Jabine & Rothwell (1970), which blocked on interviewers was the methods test just described. Again, we seem to see the areas of sampling and experimental design moving apart. Nonetheless, a later study, carried out by the Census for the Committee on National Statistics' Panel on Privacy and Confidentiality as Factors in Survey Response (1979), shows the importance of blocking on interviewers for detecting differences in response rates to different guarantees of confidentiality.

In response to an inquiry about the uses of embedding at the Census Bureau, Hansen (1984) noted:

> Although the designs of many of the test studies resemble classical experimental designs, the major objective in comparing alternatives was to estimate the difference in quality, cost, etc., between them, with an estimated sampling error, in the context of the application in mind. . . . Suppose it were believed that an interviewer could not deal adequately with more than two different versions. Then the design might look like a balanced

incomplete block design (with the interviewer being a block) in a cross-over experiment (half the interviewers who were assigned to use a given pair of alternatives starting with one of the two, and the other half starting with the other alternative). However, critical elements in the design would be the number of interviewers, and the amount of work (say, interviews) to be done by each. . . . The number of interviewers [would be] important to reduce the sampling variance of estimates of differences between procedures because of the between-interviewer component of variance arising from the correlated error within interviewers. This [would not be] simply a matter of the recovery of inter-block information as classically discussed for balanced incomplete block designs.

The experimental design perspective leads naturally to the *combination* of *local control within interviewer* to compare alternative questionnaires with the *measurement of interviewer differences* through an IPNS-like structure (Yates, 1981, pp. 110–111). This suggests the use of factorial or fractional factorial designs in randomized blocks; see, for example, Durbin & Stuart (1951), where respondents are blocked in homogeneous groups and interviewers are treated as a factor within blocks, but are also linked via an IPNS throughout the blocks. Such designs are of greatest value as part of a special investigation or a pilot survey (recall the remarks of Hansen above). Other examples of the use of experimental designs in surveys are given by Brunk & Federer (1953), Quenouille (undated) and Brewer et al. (1977).

A final illustration of the embedding of methodologies can be found in the new design for the National Assessment of Educational Progress (NAEP) survey that was adopted in 1983 by the Educational Testing Service (Messick, Beaton & Lord, 1983). Initiated in the 1960's, NAEP's primary purpose has been to measure educational competence in various subject areas (e.g. mathematics, reading and writing) primarily for three age groups (9-year olds, 13-year olds and 17-year olds) at a national level in the United States. The new design builds on that used for the first 14 years of the survey and has the following features:

(a) a deeply-stratified three-stage sampling plan, where the third stage involves randomly selecting students within a school. At stage 1, primary sampling units (PSU's) are grouped into strata by geography and community type, with an estimate of at least 1500 youths within each PSU at each of the three assessment ages. At stage 2, both public and private schools are selected within PSU's, and then at stage 3, random samples of students are selected for participation.
(b) sampling by school grade as well as by age. The meaning of grade level varies according to the age at which children enter school in each local school district, e.g. only about 70% of 9-year olds are in grade 4 and only about 70% of grade 4 are 9-year olds. Thus expanding the sample (using a dual frame) to include students in the modal grade for each age group gives a different perspective on performance level and trends.
(c) repeated school participation in successive waves of the survey, according to a rotation schedule in which 50% of the PSU's are identical in successive waves for the same subject area. The correlation within schools between waves yields sizable reductions in sampling errors for measures of change.
(d) an assignment of exercises to respondents within a particular subject area using the balanced incomplete block spiralling variant of matrix sampling.

In traditional matrix sampling, as originally used in NAEP, the exercise pool for a given age level is divided into different booklets with each respondent receiving one booklet (and thus a subset of the questions to answer), e.g. 5 booklets of 33 exercises each. The exercises appearing in one booklet do not appear in another at the same age level. The BIBD design superimposed upon matrix sampling introduces a restructured allocation of exercises. Thus in (d) above, each exercise is administered the same number of times as it would be in matrix sampling but each pair of exercises is also administered together in the same booklet a prescribed number of times. Continuing with the above example, the $5 \times 33 = 165$ exercises can be combined into 15 blocks of 11 each, and these blocks can be permuted so that each pair of blocks appears together in at least one booklet, for a total of 35 booklets. Such a design not only introduces more efficient estimates, e.g. of overall proportions responding correctly to an exercise, than was the case with matrix sampling alone, but it also allows for the estimation of other quantities of interest through the construction of a complete cross-products matrix for exercises because each exercise is located in several different booklets, each with a differing subset of the other exercises.

5 Possible causes and curses of specialization

We have maintained, with Kish (1965, p. 595) that '[t]he separation of the sample surveys from experimental designs is an accident in the recent history of science . . .' and, indeed, we embarked on this work to contribute to the accuracy of the prediction with which he continues: '. . . and it will diminish'. Here we speculate, without claiming expertise as historians, on the possible reasons for this separation.

It seems to us that although the principles of the design of randomized experiments and the design of sampling studies have a common theoretical base, the evolution of thinking in the two areas is such that the focus of each shifted into directions that, on the surface, appear to overlap less and less with the other area. In particular, in the design of randomized experiments, focus shifted toward multiple factors and complex *treatment structures* rather than more and more complex *control structures*. (Here we distinguish between control structures, e.g. blocking, Latin squares, and the treatment structure, e.g. full or partial factorial layouts, which together make up the design.) It is only when one gets to a complex design that this distinction between treatment structure and control structure arises; in a simple design (say with only one factor) one can move back and forth with ease. Since the immediate analogues of treatments in a sampling setting are the samples, and since in the simplest of worlds we select only a single sample (see § 2, above), the applicability of ideas from experimental design involving treatment structures more complex than a single factor at two levels to sampling would well have become less clear. Conversely, in sampling theory, where the theory began with simple random samples and then was elaborated with the imposition of *control* structures (e.g. stratification and clustering), real innovations arose with the consideration of whether it was advantageous to allow for unequal probabilities of selection. The concept of unequal probabilities of selection would have been one for which an experimental analogue is not immediately obvious; but see Cox (1956), who demonstrated the value of weighted randomization in designs in which a covariance adjustment is made for concomitant variation.

In recalling why the design of sample surveys took a direction so different from the design of experiments, Madow (1981) noted three additional factors:

(a) the heterogeneity, skewness, and mixture properties of the populations sampled;
(b) the large sizes of samples selected in sampling from finite populations made it possible to draw inferences that *did* depend on a probability structure imposed by the survey designer and *did not* depend on assumed probability densities;
(c) from the early work of Fisher, for example the 1925 book *Statistical Methods for Research Workers,* the simplest analysis of variance model did not permit a negative intraclass correlation coefficient, while cluster sampling as defined for finite sampling would yield a negative intraclass correlation.

For Madow, the decisive factors that turned his work on sample surveys into new directions were the negative intraclass correlation and the cost-effectiveness of unequal probability sampling.

Another difference in the purpose of random sampling and randomized experiments that could have influenced the practitioners differentially is that historically many sample surveys were designed as enumerative studies while most experiments were clearly analytic. Thus sample surveys were designed to estimate aggregates for populations, and experiments were designed to explore the causal relations between variables; see for example, Deming (1953, 1978). In fields where experimentation is sometimes difficult (including, but not limited to, parts of the social sciences), however, data derived from sampling studies must often be used for analytic purposes. The confluence would seem to raise both hopes and expectations for the transfer of methodology from the experimental realm to the sampling realm.

In addition to these differences in purposes, differences in the statistical problems faced in analysis by the two areas could well have discouraged researchers from searching for techniques in the literature of the opposite area. For example, because analyses of nonorthogonal experiments were more difficult than those based on neatly orthogonal designs, despite the pioneering work of Yates (1934) on unbalanced designs, many experiments were carefully designed to preserve orthogonality, and techniques were developed to mimic orthogonality in the analysis when failure to achieve it resulted from experimental exigencies. Surveys, on the other hand, even when used analytically, rarely achieve orthogonality. Indeed, when the sizes of subpopulations are disparate, orthogonality is not even an aim. Those techniques developed for analysis in orthogonal or almost-orthogonal designs in the experimental literature seemed, perhaps, less than applicable to the nonorthogonal analyses that survey researchers faced.

Simultaneity of inference represents another difference between sampling studies and experiments at the analysis phase. While the number of comparisons in an experiment may be large, most of them can be anticipated and planned for in advance. The number of comparisons possible in a moderately large-scale sample survey is enormous; indeed, analyses are often followed by secondary analyses from archived data. Thus there is a temptation for a survey analyst to go on fishing expeditions within the data, and while this tendency has been decried and discouraged as a capitalization on chance (see, for example, Selvin & Stuart (1966)) nevertheless survey researchers may have been discouraged from consulting literature originating in the experimental field which visualized only small numbers of comparisons and exacted a heavy statistical penalty for making a multiplicity of significance tests. The solution that is advocated in the sample survey literature, breaking the sample into random pieces and exploring on one piece to develop comparisons as hypotheses to be tested on the reserved piece, seems to have no analogue in the experimental literature.

There is also the common sense argument that because so many of the key figures in statistics in the early 20th century worked in both fields, they naturally provided rich cross-fertilization. But as the literature grew past the point where it could be easily mastered in its entirety by a newcomer to the field, and as such newcomers did indeed enter the field, specialization became necessary, and a convenient division of labour arose between those who were developing and applying techniques to sampling problems and those who were developing and applying (often similar) techniques in experimental settings. Further, the use of statistics has spread from agriculture and engineering to a diversity of applications in biomedical and social sciences and elsewhere. Each discipline in these areas seems to fasten on a particular technique for the collection of data; for example, psychologists often use experimental methods while sociologists tend to analyse sample surveys. This specialization, in turn, fosters the development of specialized and separate techniques for the different disciplines and encourages their exposition in separate textbooks.

All of these reasons for the separation of the fields implied, it seems to us, a testable hypothesis: the relatively early literature in each of the fields should show more frequent cross-citations to the other field, while the more recent literature should show far fewer cross-citations. We carried out an informal test of this hypothesis by comparing two arbitrarily chosen years (1948 with 1983) of the *Annals of Mathematical Statistics* (and the *Annals of Statistics*) and the *Journal of the American Statistical Association.* To our surprise we found no cross-citations in 1948 and only a specialized few in 1983. These later cross-citations occurred in the work of such authors as Meedan & Ghosh (1983) and Cheng & Li (1983), who are engaged in conscious programs of research to import theory and methods from one field to the other. What we did find in 1948, however, was authors such as Stephan discussing both fields in the same breath. It seems to us that the very fact of thinking of the fields as essentially synonomous accounts for the absence of the cross-citations that would flag parallels in the early years; one has to think of things as separate before one can think of them as parallel.

Thus we can make no clear conclusion about the causes of the separation. We can be more sure, however, about the penalty both fields paid for this specialization. While researchers rarely re-invent the wheel, they frequently construct less perfect vehicles for investigation than they might if they were to take full advantage of all the theory and methodology in both areas.

6 Modelling and inference

There are possible conflicts associated with the two basically different approaches to the analysis of sampling and experimental data. The reporting of information from sample surveys often takes the form of cross-classifications of frequencies, totals, means or rates. Such surveys are referred to as being *descriptive* or *enumerative* rather than *analytic* in purpose (Deming, 1953, 1978). For an enumerative survey there is general agreement in the statistical community that the reported information should consist of weighted estimates of population quantities, with the weights to be applied to individual sample cases determined by the probabilities of selection for inclusion in the sample (and possibly also by noncoverage and nonresponse rates and by ratio adjustments).

In contrast, in an analytic sample survey the primary purpose may be the comparison of sectors or subgroups of the population, often defined by multiple factors, with respect to one or more variables of interest, or some other form of comparative estimation. Statistical models often play an important role in the planning of analytic sample surveys. For example, Sedransk (1965, 1967) uses ANOVA models to illustrate how to select samples to achieve the maximum precision for the comparisons of interest for a specified total cost. He notes in particular the importance of reflecting the relevant population for inference in the planning model used for allocation. Thus the controversy arises, not in the use of models for designing surveys, but in their use for analysis; for example, see Hansen, Madow & Tepping (1983) and Smith (1985).

For the analysis of designed experiments, most modern statistical texts use a presentation based on analysis of variance (ANOVA) models with normally distributed error terms, in accord with Fisher's original development (Fisher, 1925). The later justification for ANOVA-like procedures proposed by Fisher (1935) had less to do with a belief of the appropriateness of the normal error model. Rather, Fisher constructed tests for the effect of different treatments based solely on the random assignment of treatments to experimental units, and argued that such randomization justifications were the only valid basis for inference in experimental settings (Cochran, 1978; Finney, 1964). Yates (1964) suggests that Fisher did not really regard the use of randomization tests as reasonable and he refers to a sentence inserted in the 7th edition of *The Design of*

Experiments supporting this view. Nonetheless, Fisher's development of randomization tests for experiments lies at the core of the design-based approach to inference. Later Kempthorne (1952, 1955) and others elaborated upon Fisher's arguments to show the approximate validity of ANOVA and normal-theory-based inferences for complex experimental designs.

This happy confirmatory support for model-based approaches to inference from randomization theory does not carry over immediately to the sampling context despite the many structural parallels between experimental design and sampling design. The extent to which such carryover holds is part of the controversy over the use of models for inferences in sampling settings. It is, in fact, somewhat paradoxical that Neyman, Iwaszkiewicz & Kolodziejczyk (1935), in their critique of the effectiveness of Latin square designs, used a model-based form of inference to argue against Fisher's randomization analysis whereas Neyman used a randomization analysis in his 1934 paper to argue against a model dependent approach to sampling.

Kalton (1983) provides an excellent and balanced review of the controversy surrounding the use of models in sampling practice. Hansen et al. (1983) also review aspects of the dispute with an emphasis on the randomization approach and its strengths, while Smith (1985) does the same but with emphasis on the modelling approach and its strengths. (See also the special June 1984 issue of *Survey Methodology* devoted to different accounts of issues and methods for the analysis of sampling data.) What Kalton (1983) notes is that, whatever one's views are regarding the use of models for inference in analytic surveys, models are essential to dealing with nonresponse and attrition. Often there is simply an implicit model that we can ignore the mechanism that led to the missing data. Most imputation methods have imbedded within them this simple ignorable-nonresponse model, but within strata. Methods that 'reweight' a sample to 'adjust for nonresponse' are based on a similar form of model. Hiding these modelling assumptions in the cloak of weights makes it difficult to determine the validity of weighted analyses of the sort referred to above, even for enumerative surveys. See Fienberg & Tanur (1986a) for a review of model-based approaches that pay special attention to the nonresponse mechanism.

It seems to us that once there are unequal probabilities of selection so that the sample is not self-weighting, even within strata, it is no longer clear that the modelling and design-based inference approaches should be expected to coincide. Advocates for the model-based approach to survey analysis argue that aspects of the sampling design such as critical stratification variables and cluster effects should be incorporated into the statistical model, with relevance being an empirical issue. In the best of all worlds, the formulation of the model precedes the design of the sample and thus the design reflects the critical stratification variables and cluster effects rather than the design dictating what should go into the analysis model. If the model assumes some level of homogeneity (as reflected in the error term), then this assumption needs to be critically examined using standard diagnostic approaches. From this perspective, conventional weighting (using the inverse of the 1st-order inclusion probabilities) to achieve 'representativeness' is a needless complication.

Strong support for the modelling position comes from statistical likelihood theory via arguments articulated by Smith (1983) and by Hoem (1985). Smith uses an approch due to Rubin (1976) to show that the sample selection scheme is ignorable if the selection of units depends only on prior variables which are conditioned on in the statistical model. In the Appendix we give a version of Smith's argument on when weighting can be ignored from a model-based perspective and we discuss the relevance of random selection and ignorability both from a likelihood and a Bayesian perspective. The discussion by Rao

(1975) is closely related to that in Smith's paper. In the context of parallels, Smith's argument mirrors one given by Rubin (1978) on the role of experimental randomization in guarding against data poorly balanced with respect to blocking variables. The carryover of Rubin's approach is immediate once we recall that the sampling analogues to the blocking variables are the stratification variables in the explanatory structure of the model, and that random sampling plays a role analogous to experimental randomization in making various types of sampling information ignorable. Smith & Sugden (1985) examine many of these issues regarding inferences in sample survey settings in greater detail.

While the differences between the design-based and model-based approaches to inference can be profitably minimized in many practical situations such as those involving regression models (Särndal, 1980), in others conflicts and controversies are inevitable; see, for example, Godambe (1955, 1982). One way to finesse the issue is to do the analysis both ways in order to be sure that the conclusions are the same. For example, Schirm et al. (1982), in a demographic study of contraceptive failure, note that

> fortunately, our early analysis of these data revealed that the estimates of effects derived using an unweighted sample are approximately equal to those obtained using a weighted one; therefore, in the subsequent analysis, the . . . weights are ignored.

The analyses these authors describe are ones for which the weights are irrelevant (Hoem, 1985). Yet by reporting that the results are unchanged if weighted analyses are done, they ensure that the focus is on their results and not on the controversy over the use of weights. In a similar spirit of reconciliation we would do well to keep in mind the advice of Rao (1975) in this regard:

> . . . many of the controversies could be avoided if the basic issues are clearly stated and statisticians do not insist on a monolithic structure for all problems of statistical inference. Much damage has been done by fashions and slogans in statistics introduced by theoretical statisticians who have no experience of handling live data and extracting information from them.

7 The tradition continues

In this paper we have retraced some of the history of the design and analysis of randomized experiments and sampling studies. Beginning with the work of Fisher and Neyman, we have followed some of the many intertwining and parallel paths of research in the two areas up to the present. Implicit in our discussion has been an answer to the question 'What can experts in sampling and experimental design learn from one another?' We find the concepts and constructs in the two areas to be so closely linked that it is surprising that only a few experts in sampling have already learned to draw on experimental design, and vice versa.

Yet we have also noted how research and practice in experimentation and in sampling have grown apart, and thus efforts in one area often fail to take advantage of the theory and methodology in the other. The lack of cross-references in the review papers by Cox (1984) and Smith (1984) suggests that the specialization extends even to compartmentalization within the minds and professional lives of outstanding investigators, for both these authors have been steeped in the tradition of parallels. (In fact, Cox mentions the existence of many parallels, in a single sentence, but offers no references.) To accomplish the transfer of theory and methodology between the two areas in the future, we need an understanding of the history and models of how parallels can be discovered and usefully applied. Three such models come to mind.

First, consider a sampling statistician who has a new problem to solve. Such a statistician ought to try to read the experimental design literature in order to find a parallel solution to exploit. Secondly, without an explicit practical problem, either an experimentalist or a sampling person might generate ideas for new methodological research by focusing on a feature in one of the literatures and trying to construct a parallel in the other. For example, someone might profitably ask what the sampling parallel to mixture designs might be. But from the research perspective perhaps the most intriguing model is to find new areas of activity, e.g. Monte Carlo analysis and simulation, where the methodologies from both areas can be profitably imported and used together. For example the experimental device of 'antithetic variates' and the method of 'importance sampling', a version of sampling with probabilities proportional to size, are often used to study different aspects of large Monte Carlo studies but they are rarely effectively integrated in the way that takes full advantage of the intertwining concepts and constructs in experimental and sampling design.

We are not the first to adopt the theme of parallels or to suggest these models for transfer. Just over 20 years ago Dalenius & Matérn (1964) adopted the same themes and argued for the synthesis that our models suggest. They even used the example of Monte Carlo procedures as a testing ground. Our goal is something short of a Grand Unified Theory so hotly pursued by physicists. We would simply like to ensure that the tradition begun by Fisher and Neyman continues.

Acknowledgements

This research was supported in part by the National Science Foundation under Grants No. SES-84-06952 to Carnegie-Mellon University, Grant No. SES-84-06721 to the Research Foundation of the State University of New York, and Grant No. BNS-8011494 while the first author was a John Simon Guggenheim Fellow at the Center for Advanced Study in the Behavioral Sciences. An earlier version of this paper, Fienberg & Tanur (1985), was presented at the ISI Session in Amsterdam. We have benefited greatly from materials and recollections provided by D.R. Bellhouse, R. Beran, D.R. Brillinger, C.-S. Cheng, D.R. Cox, T. Dalenius, R. Dawes, W.E. Deming, Y. Dodge, D.J. Finney, J. Gower, M.H. Hansen, A. Hedayat, R. Iachan, O. Kempthorne, L. Kish, Wm. H. Kruskal, Wm. G. Madow, E. Marks, H. Nisselson, C.R. Rao, J.N.K. Rao, J. Sedransk, T.M.F. Smith, J. Steinberg, B.J Tepping, and J. Waksberg, and from comments from the editor and referees. We have a special debt to the late Wm. G. Cochran who taught the first author several years ago about the importance of the parallels between experiments and surveys. It is the tradition that he helped to establish which we are attempting to continue.

Appendix: Model-based inferences for sampling

From a model-based perspective many authors (see, for example, Basu (1969, 1978)) have argued that, at the data analysis stage, the survey analyst need not be concerned with the sample design, and thus notions such as the sampling distribution, design-unbiasedness, and sample-based weights are irrelevant. In particular Godambe (1966) argued that the likelihood principle implies that inferences should be independent of the distribution generated by the sampling design. These statements are not quite correct, and an appropriate argument for when they are correct has been developed independently by Smith (1983) and Hoem (1985). In the Appendix, we present a version of their basic result using Smith's notation and provide some commentary on it.

Following the notation in § 2, we denote by s the sample and by $\mathbf{A}_s$ the sample selection function whose components are the indicator variables

$$A_i = \begin{cases} 1 & (i \in s), \\ 0 & (i \notin s). \end{cases} \tag{A.1}$$

Corresponding to each unit u_i in the universe $\mathbf{U}$, there is a vector of unknown values or measurements, $\mathbf{Y}_i$, and a vector of prior information, $\mathbf{Z}_i$. The sampling scheme or selection mechanism is given by the probability density function

$$f(\mathbf{A}_s \mid \mathbf{Z}, \mathbf{Y}; \boldsymbol{\phi}), \tag{A.2}$$

when $\mathbf{Z}$ is the matrix of prior information, $\mathbf{Y}$ is the matrix of measurement variables, and $\boldsymbol{\phi}$ is a possibly unknown parameter vector characterizing the selection process. Note that in this specification we allow for a dependence of the selection on the outcomes of interest, $\mathbf{Y}$. Hoem (1985) gives several examples of selection mechanisms that appear to involve forms of 'random selection' but which in fact reflect such dependence. For example, in sample surveys involving retrospective interviews the sample is usually drawn from a frame representing 'survivors', i.e. dead individuals are unavailable for sampling. Now if mortality is related to the outcome measurements in some selective fashion, the sampling mechanism will depend on $\mathbf{Y}$ as in (A.2).

Next, we introduce the superpopulation model of interest which links the measurement values $\mathbf{Y}$ to the prior information via the conditional distribution with probability density function

$$f(\mathbf{Y} \mid \mathbf{Z}; \boldsymbol{\theta}). \tag{A.3}$$

We would like to make inferences about θ. Now we can partition $\mathbf{Y}$ into two components $\mathbf{Y}_s$ and $\mathbf{Y}_{\bar{s}}$ depending on whether the units are in s or are not in s, respectively. Then the observed data consist of

$$\mathbf{d}_s = (\mathbf{Y}_s, \mathbf{A}_s), \tag{A.4}$$

an observation that can be found in the earlier literature, for example, Basu (1978), and the probability density function of the observed data is given by

$$f(d_s \mid \mathbf{Z}; \boldsymbol{\theta}, \boldsymbol{\phi}) = \int f(\mathbf{Y}_s, \mathbf{Y}_{\bar{s}} \mid \mathbf{Z}; \boldsymbol{\theta}) f(\mathbf{A}_s \mid \mathbf{Y}_s, \mathbf{Y}_{\bar{s}}, \mathbf{Z}; \boldsymbol{\phi}) d\mathbf{Y}_{\bar{s}}. \tag{A.5}$$

This is the likelihood function and it clearly depends on the probability distribution associated with the selection mechanism.

Suppose that the selection scheme in (A.2), depends on $\mathbf{Z}$ but not on $\mathbf{Y}$, i.e.,

$$f(\mathbf{A}_s \mid \mathbf{Y}, \mathbf{Z}; \boldsymbol{\phi}) = f(\mathbf{A}_s \mid \mathbf{Z}; \boldsymbol{\phi}). \tag{A.6}$$

This is where randomization enters into the argument for those who believe in model-based inferences. If a randomization design has been employed whereby all units of the universe have positive probability of selection, π_i, and these probabilities do not depend on the $\mathbf{Y}_i$'s, then (A.6) holds. It follows that

$$f(\mathbf{Y}_s, \mathbf{A}_s \mid \mathbf{Z}; \boldsymbol{\theta}, \boldsymbol{\phi}) = f(\mathbf{A}_s \mid \mathbf{Z}; \boldsymbol{\phi}) \int f(\mathbf{Y}_s, \mathbf{Y}_{\bar{s}} \mid \mathbf{Z}; \boldsymbol{\theta}) d\mathbf{Y}_{\bar{s}} = f(\mathbf{A}_s \mid \mathbf{Z}; \boldsymbol{\phi}) f(\mathbf{Y}_s \mid \mathbf{Z}; \boldsymbol{\theta}). \tag{A.7}$$

Provided that $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are distinct, the likelihood function in (A.7) separates into two components, one for $\boldsymbol{\theta}$ and one for $\boldsymbol{\phi}$, and inferences about $\boldsymbol{\theta}$ based on (A.7) will be equivalent to those based on

$$f(\mathbf{Y}_s \mid \mathbf{Z}; \boldsymbol{\theta}) = \int f(\mathbf{Y}_s, \mathbf{Y}_{\bar{s}} \mid \mathbf{Z}; \boldsymbol{\theta})\, d\mathbf{Y}_{\bar{s}}. \tag{A.8}$$

Expression (A.8) is often taken to represent the likelihood function.

For the sampling plan to be ignorable, in the sense that inferences about θ are the same from (A.7) and (A.8), we need not assume that there has been a formal use of randomization in the selection scheme. Any selection plan for which (A.6) holds will do, but the standard forms of random selection remove the need for further justification. Thus randomization plays the role of removing selection bias.

The preceding argument is extremely simple but it is applicable quite broadly and can be extended to include both selection and allocation mechanisms, the latter being relevant for experiments, as well as allowing a separate look at nonresponse. Several additional comments and observations may be of value as follows.

(i) The form of the superpopulation model in (A.3) hides a multiplicity of complexities that possibly relate to the sample design. For example, suppose that the model takes the form

$$\mathbf{Y} = \boldsymbol{\theta}_1 \mathbf{Z} + \boldsymbol{\varepsilon}. \tag{A.9}$$

Then, $\mathbf{Z}$ would be likely to reflect important stratification variables that would also go into the design and the covariance structure of $\boldsymbol{\varepsilon}$ would reflect dependencies among units due to clustering; see the related discussion in § 6.

(ii) The result that inferences from (A.7) and (A.8) are equivalent *does not depend* on proper specification of the superpopulation model (Hoem, 1985). It is for this reason we have a difficult time understanding the robustness arguments raised by Hansen et al. (1983), Kalton (1983) and others. This comment should not be interpreted as implying that model misspecification is not an issue. Rather we would argue that it is the same issue in a sampling context as in any other statistical problem.

(iii) Simple sampling estimation problems, such as heterogeneity across strata, are easily represented within the general superpopulation specification of (A.3) and can result in likelihood-based estimates of quantities such as population means that are similar to the traditional weighted design-based estimates. Post-stratification models also fit easily within this broad framework.

(iv) We need to distinguish between sampling plans that are *informative* and those that are *ignorable*. Condition (A.3) corresponds to Rubin's technical notion of ignorability. Survival as a condition for observation in a sample makes a plan informative, but if survival is nonselective then the sampling plan is still ignorable; see Hoem (1985) for a more elaborate discussion of this issue.

(v) The argument above is presented from a likelihood perspective, but it is virtually identical to one relevant to the Bayesian. The separability of the likelihood in (A.7) is not sufficient to allow the selection mechanism to be ignored. In addition, the Bayesian requires that $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ be a priori independent, although, as Smith (1983) notes, this allows a somewhat weaker condition in the relationship between $\mathbf{Y}$ and $\mathbf{A}_s$; see also Rubin (1976).

(vi) From the perspective of comment (v), the easiest way for a Bayesian to justify basing inferences solely on (A.8) and the prior distribution for $\boldsymbol{\theta}$ is to use careful randomization in the selection process; see Rubin's (1978) argument for why a Bayesian should randomize in an experimental context.

(vii) The Bayesian who chooses to use randomization gains the simplicity of the likelihood in (A.8) at a price. An informative and nonignorable sample plan can produce a posterior distribution for $\boldsymbol{\theta}$ that is tighter than that resulting from (A.8). This loss of information can be substantial if n is small, but we conjecture that it tends to zero in some stochastic sense as n tends to N, for large values of N.

References

Avadhani, M.S. & Sukhatme, B.V. (1973). Controlled sampling with equal probabilities and without replacement. *Int. Statist. Rev.* **41,** 175–182.

Bailey, R.A. (1985). Restricted randomization versus blocking. *Int. Statist. Rev.* **53,** 171–182.

Bartlett, M.S. (1978). Fisher, R.A. In *The International Encyclopedia of Statisitcs,* Ed. W.H. Kruskal and J.M. Tanur, **1,** pp. 352–358. New York: Free Press.

Basu, D. (1969). Role of sufficiency and the likelihood principles in sample survey theory. *Sankhyā,* **31,** 441–454.

Basu, D. (1978). Relevance of randomization in data analysis. In *Survey Sampling and Measurement,* Ed. N.K. Namboodiri, pp. 267–292. New York: Academic Press.

Bellhouse, D.R. (1977). Some optimal designs for sampling in two dimensions. *Biometrika* **64,** 605–611.

Bellhouse, D.R. (1984a). Optimal randomization for experiments in which autocorrelation is present. *Biometrika* **71,** 155–160.

Bellhouse, D.R. (1984b). A review of optimal designs in survey sampling. *Can. J. Statist.* **12,** 53–65.

Bellhouse, D.R. (1986). Randomization in the analysis of covariance. *Biometrika* **73,** 207–211.

Box, J.F. (1978). *R.A. Fisher, The Life of a Scientist.* New York: Wiley.

Brewer, K.R.W., Foreman, E.K., Mellor, R.W. & Trewin, D.J. (1977). Use of experimental design and population modelling in survey sampling. *Bull. Int. Statist. Inst.* **3,** 173–190.

Brunk, M.E. & Federer, W.T. (1953). Experimental designs and probability sampling in marketing research. *J. Am. Statist. Assoc.* **48,** 440–452.

Causey, B.D., Cox, L.W. & Ernst, L.R. (1985). Applications of transportation theory to statistical problems. *J. Am. Statist. Assoc.* **80,** 903–909.

Chakrabarti, M.C. (1963). On the use of incidence matrices of designs in sampling from finite populations. *J. Ind. Statist. Assoc.* **1,** 78–85.

Cheng, C.-S. & Li, K.-C. (1983). A minimax approach to sample surveys. *Ann. Statist.* **11,** 552–563.

Cheng, C.-S. & Li, K.-C. (1985). Optimality criteria in survey sampling. Unpublished manuscript.

Cochran, W.G. (1977). *Sampling Techniques,* 3rd ed. New York: Wiley.

Cochran, W.G. (1978). Experimental design: the design of experiments. In *The International Encyclopedia of Statistics,* Ed. W.H. Kruskal and J.M. Tanur, **1,** pp. 285–294. New York: Free Press.

Cochran, W.G. & Cox, G.M. (1957). *Experimental Designs,* 2nd. ed. New York: Wiley.

Cochran, W.G. & Watson, D.J. (1936). An experiment on observer's bias in the selection of shoot heights. *Empir. J. Exper. Agric.* **4,** 69–76.

Cornfield, J. & Tukey, J.W. (1956). The average values of mean squares in factorials. *Ann. Math. Statist.* **27,** 907–949.

Cox, D.R. (1956). A note on weighted randomization. *Ann. Math. Statist.* **27,** 1144–1151.

Cox, D.R. (1958). *Planning of Experiments.* New York: Wiley.

Cox, D.R. (1984). Present position and potential developments: some personal views, design of experiments and regression. *J. R. Statist. Soc.* A **147,** 306–315.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *Am. Psychol.* **12,** 671–684.

Dalenius, T. & Matérn, B. (1964). Is there a need for a unified theory of random experiments? *Metrika* **8,** 235–247.

Deming, W.E. (1953). On the distinction between enumerative and analytic surveys. *J. Am. Statist. Assoc.* **48,** 244–255.

Deming, W.E. (1978). Sample surveys: the field. In *The International Encyclopedia of Statistics,* Ed. W.H. Kruskal and J.M. Tanur, **2,** pp. 867–885. New York: Free Press.

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

Fienberg, S.E. (1971). Randomization and social affairs: the 1970 draft lottery. *Science* **171,** 255–261.

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

Fienberg, S.E. & Tanur, J.M. (1983). Reuniting the twain: remarking on the parallels between sample surveys and randomized experiments. Paper presented at International Statistical Institute Meetings, Madrid, Spain.

Fienberg, S.E. & Tanur, J.M. (1985). A long and honorable tradition: intertwining concepts and constructs in experimental design and sample surveys. *Bull. Int. Statist. Inst.,* 44th Session, Book II, 10.1-1–10.1-18.

Fienberg, S.E. & Tanur, J.M. (1986a). The design and analysis of longitudinal surveys: controversies and issues of cost and continuity. In *Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits,* Ed. R. Boruch and R. Pearson, pp. 60–93. New York: Springer-Verlag.

Fienberg, S.E. & Tanur, J.M. (1986b). From the inside out and the outside in: combining experimental and sampling structures. Technical Report No. 373, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Finney, D.J. (1964). Sir Ronald Fisher's contributions to biometrics statistics. *Biometrics* **20,** 322–329.

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

Fisher, R.A. (1926). The arrangement of field experiments. *J. Ministry Agric. G.B.* **33,** 503–513.

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

Foody, W. & Hedayat, A. (1977). On theory and applications of BIB designs with repeated blocks. *Ann. Statist.* **5,** 933–945.
Frankel, L.R. & Stock, J.S. (1942). On the sample survey of unemployment. *J. Am. Statist. Assoc.* **37,** 77–80.
Godambe, V.P. (1955). A unified theory of sampling from finite populations. *J. R. Statist. Soc.* B **17,** 269–278.
Godambe, V.P. (1966). A new approach to sampling from finite populations. *J. R. Statist. Soc.* B **28,** 310–328.
Godambe, V.P. (1982). Estimation in survey sampling: robustness and optimality (with discussion). *J. Am. Statist. Assoc.* **77,** 393–406.
Goodman, R. & Kish, L. (1950). Controlled section—a technique in probability sampling. *J. Am. Statist. Assoc.* **45,** 350–372.
Grundy, P.M. & Healy, M.J.R. (1950). Restricted randomization and quasi-Latin squares. *J. R. Statist. Soc.* B **12,** 286–291.
Hansen, M.H. (1984). Personal communication.
Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953a). *Sample Survey Methods and Theory,* **1,** New York: Wiley.
Hansen, M.H., Hurwitz, W.N. & Madow, W.G. (1953b). *Sample Survey Methods and Theory,* **2**. New York: Wiley.
Hansen, M.H., Hurwitz, W.N., Marks, E.S. & Mauldin, W.P. (1951). Response errors in surveys. *J. Am. Statist. Assoc.* **46,** 147–190.
Hansen, M.H., Madow, W.G. & Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion). *J. Am. Statist. Assoc.* **78,** 776–793.
Hartley, H.O. & Rao, J.N.K. (1978). Estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurement,* Ed. N.K. Namboodiri, pp. 35–43. New York: Academic Press.
Hedayat, A. (1979). Sampling designs with reduced support sizes. In *Optimizing Methods in Statistics,* Ed. J. Rustagi, pp. 273–288. New York: Academic Press.
Hoem, J.M. (1985). Weighting, misclassification, and other issues in the analysis of survey samples of life histories. In *Longitudinal Analysis of Labor Market Data,* Ed. J.J. Heckman and B. Singer, pp. 249–293. Cambridge University Press.
Holschuh, N. (1980). Randomization and design: I. In *R.A. Fisher: An Appreciation,* Ed. S.E. Fienberg and D.V. Hinkley, pp. 35–45. New York: Springer-Verlag.
Iachan, R. & Jones, B. (1984). Rotation sampling designs, I: construction, classification, and structure. Unpublished manuscript.
Jabine, T.B. & Rothwell, N.D. (1970). Split-panel tests of census and survey questionnaires. *Proc. Social Statist. Sect.,* Am. Statist. Assoc., pp. 4–13.
Kalton, G. (1983). Models in the practice of survey sampling. *Int. Statist. Rev.* **51,** 175–188.
Keifer, J. (1961). Optimum designs in regression problems. II. *Ann. Math. Statist.* **32,** 298–325.
Kempthorne, O. (1952). *The Design and Analysis of Experiments.* New York: Wiley.
Kempthorne, O. (1955). The randomization theory of experimental inference. *J. Am. Statist. Assoc.* **50,** 946–967.
Kish, L. (1959). Some statistical problems in research design. *Am. Sociol. Rev.* **24,** 328–338.
Kish, L. (1965). *Survey Sampling.* New York: Wiley.
Kish, L. & Frankel, M.R. (1974). Inference from complex samples (with discussion). *J. R. Statist. Soc.* B **36,** 1–37.
Kruskal, W.H. & Mosteller, F. (1980). Representative sampling IV: the history of the concept in statistics, 1815–1939. *Int. Statist. Rev.* **48,** 169–195.
Madow, W.G. (1981). Personal communication.
Mahalanobis, P.C. (1944). On large-scale sample surveys. *Phil. Trans. R. Soc. Lond.* B **231,** 329–451.
Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *J. R. Statist. Soc.* **109,** 325–378.
Meeden, G. & Ghosh, M. (1983). Choosing between experiments: applications to finite population sampling. *Ann. Statist.* **11,** 296–305.
Messick, S., Beaton, A. & Lord, F. (1983). *National Assessment of Educational Progress Reconsidered: A New Design for a New Era.* NAEP Report 83-1, Princeton, NJ.
Morris, C. (1979). A finite selection model for experimental design of the health insurance study. *J. Econometrics* **11,** 43–61.
Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *J. R. Statist. Soc.* A **109,** 558–606.
Neyman, J., Iwaszkiewicz, K. & Kolodziejczyk, St. (1935). Statistical problems in agricultural experimentation. *Suppl. J. R. Statist. Soc.* **2,** 107–180.
Panel on Privacy and Confidentiality as Factors in Survey Response, Committee on National Statistics (1979). *Privacy and Confidentiality as Factors in Survey Research.* Washington, D.C.: National Academy of Sciences.
Platek, R., Rao, J.N.K. Särndal, C.E. & Singh, M.B. (1986). *Small Area Statistics: An International Symposium.* New York: Wiley.
Quenouille, M.H. (undated). Combined use of experiments and sample surveys. Unpublished manuscript.
Raghavarao, D. & Federer, W.T. (1979). Block total response as an alternative to the randomized response method in surveys. *J. R. Statist. Soc.* B **41,** 40–45.
Rao, C.R. (1975). Some problems of sample surveys. *Adv. Appl. Prob.* **7** (Suppl.) , 50–61.

Rao, J.N.K., Hartley, H.O. & Cochran, W.G. (1962). A simple procedure of unequal probability sampling without replacement. *J. R. Statist. Soc.* B **24,** 482–491.
Reid, C. (1982). *Neyman From Life.* New York: Springer-Verlag.
Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.
Rubin, D.B. (1978). Bayesian inference for causal effects: the role of randomization. *Ann. Statist.* **6,** 34–58.
Särndal, C.-E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* **67,** 639–650.
Schirm, A.L., Trussell, J., Menken, J. & Grady, W.R. (1982). Contraceptive failure in the United States: the impact of social, economic, and demographic factors. *Family Plan. Persp.* **14,** 68–75.
Sedransk, J. (1965). Analytical surveys with cluster sampling. *J. R. Statist. Soc.* B **27,** 264–278.
Sedransk, J. (1967). Designing some multi-factor analytical studies. *J. Am. Statist. Assoc.* **62,** 1121–1139.
Selvin, H.C. & Stuart, A. (1966). Data-dredging procedures in survey analysis. *Am. Statist.* **20,** 20–23.
Seng, Y.P. (1951). Historical survey of the development of sampling theories and practice. *J. R. Statist. Soc.* A **114,** 214–231.
Singh, R. & Raghavarao, D. (1975). Application of linked block designs in successive sampling. In *Applied Statistics,* Ed. R.P. Gupta, pp. 301–309. Amsterdam: North Holland.
Smith, T.M.F. (1983). On the validity of inferences from non-random samples. *J. R. Statist. Soc.* A **146,** 394–403.
Smith, T.M.F. (1984). Present position and potential developments: some personal views, sample surveys. *J. R. Statist. Soc.* A **147,** 208–221.
Smith, T.M.F. & Sugden, R.A. (1985). Inference and the ignorability of selection for experiments and surveys. *Bull. Int. Statist. Inst.,* 44th Session, Book II: 10.2-1–10.2-12.
Srivastava, J. (1985). On a general theory of sampling using experimental design. *Bull. Int. Statist. Inst.,* 44th Session, Book II: 10.3-1–10.3-16.
Stephan, F.F. (1948). History of the uses of modern sampling procedures. *J. Am. Statist. Assoc.* **43,** 12–39.
Tepping, B.J., Hurwitz, W.N. & Deming, W.E. (1943). On the efficiency of deep stratification in block sampling. *J. Am. Statist. Assoc.* **38,** 93–100.
Tchuprov, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron* **2,** 646–680.
Waksberg, J. & Pearl, R.B. (1965). New methodological research on labor force measurement. *Proc. Social Statist. Sect.,* Am. Statist. Assoc., pp. 227–237.
Wilk, M.B. & Kempthorne, O. (1956). Some aspects of the analysis of factorial experiments in a completely randomized design. *Ann. Math. Statist.* **27,** 950–985.
Wynn, H.P. (1977). Convex sets of finite population plans. *Ann. Statist.* **5,** 414–418.
Yates, F. (1934). The analysis of multiple classifications with unequal numbers in the different classes. *J. Am. Statist. Assoc.* **29,** 52–66.
Yates, F. (1948). Discussion of a paper by F.J. Anscombe. *J. R. Statist. Soc.* A **111,** 204–205.
Yates, F. (1964). Sir Ronald Fisher and the design of experiments. *Biometrics,* **20,** 307–321.
Yates, F. (1981). *Sampling Methods for Censuses and Surveys,* 4th ed. New York: Macmillan.
Yates, F. (1985). Book review of "W.G. Cochran's Impact on Statistics", Ed. P.S.R.S. Rao and J. Sedransk. *Biometrics* **41,** 591–592.
Yates, F. & Zacopancy, I. (1935). The estimation of the efficiency of sampling with special reference to sampling for yields in cereal experiments. *J. Agric. Sci.* **25,** 545–577.
Youden, W.J. (1964). Answer to "Query 1. Inadmissible random assignments". *Technometrics* **6,** 103–104.
Youden, W.J. (1972). Randomization and experimentation. *Technometrics* **14,** 13–22.
Zarkovich, S.S. (1956). Note on the history of sampling methods in Russia. *J. R. Statist. Soc.* A **119,** 336–338.
Zarkovich, S.S. (1962). A supplement to "Note on the history of sampling methods in Russia". *J. R. Statist. Soc.* A **125,** 580–582.

Résumé

Les plans et l'analyse des expériences au hasard et des enquêtes par sondage aléatoire sont suivis à partir du travail de Fisher, Neyman, et Tchuprov dans les années 1920 et 1930, quoique des précurseurs apparaissent beaucoup plus tôt. Cet étude examine quelques développements qui découlent de leur travail, soulignant les parallèles entre les deux méthodologies. Après un examen des parallèles conceptuels fondamentaux entre les plans d'expériences et ceux de sondages, nous passons en revue une nouvelle classe de parallèles liant les formes restreintes de sondage à la litterature sur les plans d'expérience portant sur les structures de traitement, telles que le BIBD. Ces parallèles entre les deux domaines nous conduisent naturellement à discuter la facon d'incorporer les expériences aux sondages, et vice versa. Les auteurs méditent sur les raisons pour la séparation de ces deux domaines, et résument les controverses parallèles entre les deux modes d'inférence, celui fondé sur les plans et celui fondé sur les modèles, qui sont utilisés, également dans la littérature sur les plans expérimentaux et sur les enquêtes par sondages. En résumé, l'étude propose comment des idées et explications nouvelles peuvent paraître à l'avenir et enrichir la pratique future.

[*Received July* 1985, *revised July* 1986]