The Annals of Applied Statistics
2018, Vol. 12, No. 2, 1157–1179
https://doi.org/10.1214/18-AOAS1184

THE INTERLOCKING WORLD OF SURVEYS AND EXPERIMENTS

BY STEPHEN E. FIENBERG AND JUDITH M. TANUR*

*Stony Brook University**

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 [*Bull. Int. Stat. Inst.* **51** (1985) Art. ID 10.1; *Int. Stat. Rev.* **55** (1987) 75–96]). 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 trace some of parallels between sampling theory and theory of experimental design. We then explore 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 toward the improvement of practice.

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 1920s and 1930s [see the related discussions in the biographies by Box (1978) and by Reid (1982), resp.] 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 and Mosteller (1980)]. Smith and 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 70 years closer to the events, pointed out a lack of communication

Received May 2018.

Key words and phrases. External validity, internal validity, interviewer effects, randomized experiments, sample surveys, control, experimental design, embedding, randomization, sampling design.

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), page 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 to encourage cross-fertilization. It is to further encourage such cross-fertilization that we have embarked on the present research.

In Section 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. 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 Section 3, 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 Section 4, we consider an elaboration of embedding using variance component models. Then in Section 5, we summarize the issues of inference to larger populations, distinguishing between internal and external validity and considering the contributions made by the movement to study cognitive aspects of surveys. In Section 6 we explore in detail issues of inference in a specific instance of an elaborate experiment embedded in a survey. In Section 7, we note the frequent lack of follow through in taking account of design features in embedded experiments when researchers approach analysis and speculate on whether this is part of the toll taken by the growth of the field of statistics that separates specialists in experimental design from those in sample surveys. Section 8 is a brief note on the genesis of this paper.

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, $\boldsymbol{U} = \{u_1, u_2, \ldots, u_N\}$, and a sample selection function $\boldsymbol{A_s} = (A_1, A_2, \ldots, A_2)$, 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, $\boldsymbol{A}_s$, 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 non-selection. If T_i contains n members, then experimental randomization and simple random sampling both take each of the

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

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

$$\frac{1}{\binom{N}{n}}.$$

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 and Madow (1953a, 1953b), 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, that is, 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 and Cox (1957), pages 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 favored 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), pages 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), Chapter V, Cochran and Cox (1957), pages 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), pages 228ff.] or "deep" stratification [Frankel and Stock (1942),

Tepping, Hurwitz and Deming (1943), Hansen, Hurwitz and Madow (1953a), pages 480ff., Kish (1965), pages 488–495].

At a more refined level, the convenience of split-plot designs [Cochran and Cox (1957), pages 293ff.] is echoed in the analogous sampling technique, cluster sampling [Hansen, Hurwitz and Madow (1953b), Chapter 6, Cochran (1977), pages 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 and Cox (1957), pages 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), pages 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 are rare, primarily because of the heterogeneity among the units of the typical sampling population. We note, however, two Model II approaches in the sampling literature. The conceptualization of models for total survey error can take the component of variation due to interviewer as a random effect [Hansen et al. (1951), Hartley and Rao (1978)].

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. When we looked at sampling texts, we found a parallel neglect, but with some more

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

exceptions. Cochran (1977) uses the analysis of variance structure throughout as a summary device, while Hansen, Hurwitz and Madow (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 Section 3.3.

3. Embedding experiments in surveys. We can discern three main purposes for embedding experiments in surveys:

(i) to compare alternative aspects of survey methodology (questionnaires, training methods, collection methods) whether in pilot surveys, in methods test panels, or in ongoing surveys;
(ii) to make comparisons of substantive (rather than solely methodological) interest;
(iii) to explore the components of response variation and the validity of 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, Wood (A. J.) Research Corporation (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 is not usually built into the analysis; it should be.

For example, 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 a general desegregation item after a federal school intervention item and the other half were asked the questions in reverse order. The investigators found that the percentage of respondents endorsing desegregation dropped from 61.4% to 38.9% when the general desegregation question was preceded 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), page 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, Hurwitz and Madow (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), pages 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 2×3 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 surveys that are carried out in person. (While in-person interviewing has become increasingly rare in developed countries, it is still much used elsewhere. And in the developing world funding may often be scarce.)

As an example of such compromises, Fellegi (1964) combined partial IPNS and re-enumeration 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 re-enumeration 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 re-enumeration permits the estimation of more of the parameters in the responses error model than would be possible with either method alone.

The widespread application of telephone interviewing in many large-scale surveys presents the opportunity to return to the original conception of IPNS. 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)].

Of course, the rise of internet surveys which eliminate the interviewer altogether, makes a good deal of this discussion of interviewer effects moot, but many of the same ideas can be applied to variations in questionnaire format and its interactions with respondent characteristics by taking advantage of the computer's capability of being programmed to change the questionnaire on the fly according to respondents' reported demographic and other characteristics.

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 sampling [see Kish (1965), pages 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 many 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 $3^3 \times 4^2$ 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 organizations 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 (page 184):

> "Although highly elaborated 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" (page 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' 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. The issues of assigning the correct questionnaire variant to the appropriate respondent are much more tractable in an age of computer assisted interviewing.

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 component models. A broad area of applicability of experimental ideas within surveys is for the modeling and estimation of nonsampling errors using a random-effects ANOVA model. Pioneering work originated in the U.S. Census Bureau [e.g., Hansen et al. (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 modeling 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 modeling ideas. Let Y_{jt} be responses at time t for units $j = 1, 2, \ldots, n$ 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 $\overline{Y}$ to estimate Z, a "true" population quantity. Then we can decompose the deviation of $\overline{Y_t}$ from Z into three basic components:

$$\overline{Y_t} - Z = (\overline{Y_t} - \overline{\mu}_s) + (\overline{\mu}_s - \overline{\mu}) + (\overline{\mu} - Z),$$

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

$$\boldsymbol{E}(\overline{Y_t} - \overline{\mu}_s)^2 = \frac{\sigma^2}{n}\{1 + (n-1)\rho\},$$

where σ^2 is the variance of Y_{jt} over t, and ρ 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 contributions to overall variability, but can also 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. (Interviewer error may change again when the 2020 Census initiates on-line responding.) 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 re-interview 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 re-interviews 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 [Aronson, Brewer and Carlsmith (1985), Campbell (1957), Campbell and Stanley (1963), Cook and Campbell (1979)]. 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 reasonably 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. Another approach to establishing external validity would be to embed an experiment in a survey administered to a random sample of the general population. (These issues have recently be addressed under the term "Generalizability Bias" in the literature that attempts to combine results from randomized controlled trials in which careful selection criteria combine with randomization to insure internal validity, with those from careful observational studies with diverse populations to insure external validity. See, e.g., [Greenhouse et al. (2008), Kaizar (2011)].)

The move from an experimental population to a target population typically involves substantial resources not possessed by individual experimenters. 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.

A movement in cognitive psychology attempting to generalize laboratory findings to larger populations through the use of large-scale surveys arose in the 1980s [e.g., see Fienberg, Loftus and Tanur (1985) and Jabine et al. (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).

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. The embedding of statistically designed experiments within sample surveys raises issues of inference that have rarely been discussed in published sources. Despite the formal parallels in structure, there is a fundamental inferential distinction between experimental and survey contexts. Randomized statistical experiments are designed to ensure internal validity. On the other hand, sample surveys use probability sampling to ensure that results will have external validity. We have been able to discern at least three possible perspectives for statistical inference in embedded experiments:

(1) One can use the standard experiment paradigm, which relies largely on internal validity based on randomization and local control (e.g., the device of blocking) and on the assumption that the unique effects of experimental units and the

treatments effects can be expressed in a simple additive form, without interaction [Fisher (1935)]. Then inference focuses on within-experiment treatment differences.

(2) One can use the standard sampling paradigm, which, for a two-treatment experiment embedded in a survey, relies largely on external validity and generalizes the observation for each of the treatments to separate but paired populations of values. Each unit or individual in the original population from which the sample was drawn is conceived to have a pair of values, one for each treatment. But only one of these is observable, depending on which treatment is given. Then the inferences focus on the mean difference or the difference in the means of the two populations.

(3) One can conceptualize a population of experiments, of which the present embedded experiment is a unit or a sample of units, and thus capitalize on the internal validity created by the design of the present embedded experiment as well as the external validity created by the generalization from the present experiment to the conceptual population of experiments. Then inferences focus on treatment differences in a broader context than simply the present embedded experiment.

Because these three approaches focus on the same experimentally observed quantities but deal with possible inferences differently, they can potentially lead to different conclusions.

Consider, for example, an experiment to compare four different versions of a questionnaire on household income, with clusters that are part of a multistage area probability sampling design where each interviewer is assigned a cluster of four households to survey. Within a cluster, the four versions of the question are randomly assigned to households. The key response variable of interest is "reported household income" in dollars, typically transformed to a logarithmic scale. We have a randomized block design embedded in the clusters of a complex sample survey design.

(a) In the first inference approach, we use a randomization analysis for the randomized block design [Fisher (1935)] or an analysis of variance (ANOVA) model with fixed effects for both interviewers and questions, and a normally distributed error term. This analysis holds the survey design as fixed and focuses internally within clusters or interviewers on the differences in effects for the questions, thereby adjusting for the differential effects of interviewers.

(b) In the second inference approach, we divide the data into four subsets corresponding to the four versions of the question. We would then treat each subset as a sample from a population, where the sampling design is the same as that for the entire survey, but without the final stage of clustering. In each, we would estimate the average household income of the population and the corresponding standard error. Finally, we would compare the estimated population averages (although to do so properly we would need some estimate of the correlations among the four estimates induced by the within-cluster intraclass correlation). This is the

proper analysis for a standard split-ballot experiment, but more typically survey researchers ignore the correlations among the estimates. The inference here is external to the experiment and relies on the probability mechanism used to generate the sample. There is no natural way here to adjust for interviewer effects while still retaining an inference mechanism tied solely to the sample-selection probability mechanism. To deal with interviewers and their effects here, we need to consider them to be randomly selected from a fixed population of interviewers. This would then lead to something equivalent to the third approach.

(c) For the third inference approach, we have a sample of size one from a superpopulation of embedded randomized block experiments. One way to handle the inference problem is to treat the interviewers as a sample from a population of interviewers; this leads to a mixed-effects ANOVA model with interviewer effects treated as a random component and question effects treated as fixed components. [For general approaches to mixed-effects ANOVA models, see Wilk and Kempthorne (1955, 1956), Scheffé (1959). For the use of such models in the survey context, see Hartley and Rao (1978).] The formal analysis of the model here is related to, but different from, the one used in the first approach.

What is going on in this mixed-effects ANOVA model is a generalization of the treatment effect differences to the superpopulation of experiments from the present embedded experiment. The way we achieve this generalization is through representation of the interviewers as having been drawn from a superpopulation of interviewers corresponding to the conceptualized superpopulation of experiments. Thus the distinction between the first and third approaches is not simply one involving the difference between fixed and random effects in an ANOVA model but more importantly involves the level of applicability of the treatment effects.

What differences might we expect among the inference associated with the use of the three approaches in an actual experiment? If there really are differences among the interviewers, then the second approach may differ appreciably from the other two and thus would be wrong. The third approach differs from the first approach primarily through the inclusion of an extra component of variation associated with the estimated treatment effects corresponding to the "interviewer $\times$ treatment" interaction [see Scheffé (1959) for a detailed exposition of estimation in mixed-effects ANOVA models and for the related variance formulae]. Thus in the third approach an estimated difference in treatment effects will appear to be less precise than in the first approach. This is as it should be, because we need to pay an extra amount for the ability to generalize beyond the embedded experiment at hand. As a consequence, the mixed-effects model should yield "statistically significant" differences less frequently than the fixed-effects approach. The choice between the first and third approaches must depend on the intended applicability of the results.

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. 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 as 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 $\times$ 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 modeling, 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 24 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 $\times$ 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), page 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. 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. In an earlier publication, we did several illustrative empirical analyses to shed light on any differences in substantive conclusions stemming from the different perspectives [Fienberg and Tanur (1989)]. We concluded that although sometimes our three modes of inference agreed on substantive findings, there were times when, perhaps unpredictably, they did not.

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.

8. A final note from JMT. This paper is mostly a compendium of previously published work that Steve and I produced over the years. Some years before his death, Steve and I agreed to create this compendium for an edited volume [Lavrakas et al. (2018)] on the embedding of experiments in surveys. Once I had made a rough synthesis of the material from various papers we had published over the years, Steve undertook to update the work with new references he had been collecting and new ideas he had been hatching. As usual, he was producing more than any three normal humans could manage, and despite his repeated assurances that this work was at the top of his "to do" pile, he never had a chance to turn to it before his sudden decline and, despite all expectations, unbelievable death. So we are unable to present the updating he would have produced, but hope that nevertheless this compendium will be useful as a brief summary of some thinking that was important to Steve and perhaps to the larger statistical community.

REFERENCES

ARONSON, E., BREWER, M. and CARLSMITH, J. M. (1985). Experimentation in social psychology. In *Handbook of Social Psychology*, *Vol. I*, 3rd ed. (G. Lindzey and E. Aronson, eds.). Random House, New York.

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

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

BARTLETT, M. S. (1978). Fisher, R. A. In *International Encyclopedia of Statisitcs* (W. H. Kruskal and J. M. Tanur, eds.) 352–358. Free Press, New York.

BOX, J. F. (1978). *R. A. Fisher*: *The Life of a Scientist*. Wiley, New York. MR0500579

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. In *Handbook of Research on Teaching* (N. L. Gage, ed.) 171–246. Rand McNally, Chicago, IL.

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

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York. MR0474575

COCHRAN, W. G. and COX, G. M. (1957). *Experimental Designs*, 2nd ed. Wiley, New York; Chapman & Hall, London. MR0085682

COOK, T. D. and CAMPBELL, D. P. (1979). *Quasi-Experiments*: *Design and Analysis Issues for Field Settings*. Rand McNally, Skokie, IL.

WOOD (A. J.) RESEARCH CORPORATION (1959). *Woodchips*, 4, No. 1.

COX, D. R. (1958). *Planning of Experiments*. Wiley, New York; Chapman & Hall, London. MR0095561

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. E. (1984). Moving survey problems into the cognitive psychology laboratory. In *Proceedings of the American Statistical Association Section on Survey Research Methods* 19–21. Amer. Statist. Assoc., Washington, DC.

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

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

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

FIENBERG, S. E., LOFTUS, E. E. and TANUR, J. M. (1985). Cognitive aspects of health survey methodology: An overview. *Milbank Mem. Fund Q.* **63** 547–564.

FIENBERG, S. E., SINGER, B. and TANUR, J. M., (1985). Large scale social experimentation in the U.S.A. In *A Celebration of Statistics* (A. C. Atkinson and S. E. Fienberg, eds.) 287–326. Springer, New York.

FIENBERG, S. E. and TANUR, J. M. (1985). A long and honorable tradition: Intertwining concepts and constructs in experimental design and sample surveys. *Bull. Int. Stat. Inst.* **51** Art. ID 10.1. MR0886247

FIENBERG, S. E. and TANUR, J. M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *Int. Stat. Rev.* **55** 75–96. MR0962943

FIENBERG, S. E. and TANUR, J. M. (1989). Combining cognitive and statistical approaches to survey design. *Science* **243** 1017–1022. MR0986238

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

FISHER, R. A. (1926). The arrangement of field experiments. *J. Minist. Agric.* **33** 503–513.

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

FRANKEL, L. R. and STOCK, J. S. (1942). On the sample survey of unemployment. *J. Amer. Statist. Assoc.* **37** 77–80.

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

GREENHOUSE, J. B., KAIZAR, E. E., KELLEHER, K., SELTMAN, H. and GARDNER, W. (2008). Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users. *Stat. Med.* **27** 1801–1813. MR2420346

GROVES, R. M. and MAGILAVY, L. J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opin. Q.* **50** 251–266.

HANSEN, M. H., HURWITZ, W. N. and BERSHAD, M. A. (1961). Measurement errors in censuses and surveys. *Bull. Int. Stat. Inst.* **38** 359–374.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953a). *Sample Survey Methods and Theory*, *Vol. I*: *Methods and Applications*. Wiley, New York; Chapman & Hall, London. MR0058171

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953b). *Sample Survey Methods and Theory*, *Vol. II*: *Theory*. Wiley, New York; Chapman & Hall, London. MR0058172

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.

HARTLEY, H. O. and RAO, J. N. K. (1978). Estimation of nonsampling variance components in sample surveys. In *Survey Sampling and Measurement* (N. K. Namboodiri, ed.) 35–43. Academic Press, New York.

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

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, DC.

KAIZAR, E. E. (2011). Estimating treatment effect via simple cross design synthesis. *Stat*. *Med*. **30** 2986–3009. MR2851395

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*. Wiley, New York; Chapman & Hall, London. MR0045368

KEMPTHORNE, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist*. *Assoc*. **50** 946–967. MR0071696

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

KISH, L. (1965). *Survey Sampling*. Wiley, New York.

KISH, L. and FRANKEL, M. R. (1974). Inference from complex samples. *J. Roy. Statist*. *Soc*. *Ser. B* **36** 1–37. MR0365812

KRUSKAL, W. and MOSTELLER, F. (1980). Representative sampling. IV. The history of the concept in statistics, 1895–1939. *Int*. *Stat*. *Rev*. **48** 169–195. MR0586104

LAVRAKAS, P. J., TRAUGOTT, M., KENNEDY, C., DE LEEUW, E., HOLBROOK, A. and WEST, B., eds. (2018). *Experimental Methods in Survey Research*: *Techniques That Combine Random Assignment with Random Probability Sampling*. Wiley, New York. In press.

LOFTUS, E. E. and FATHI, D. (1985). Retrieving multiple autobiographical memories. *Social Cogn*. **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. In *International Encyclopedia of Statisitcs*, *Vol*. *I* (Kruskal, W. H. and Tanur, J. M., eds.) 208–229. Free Press, New York.

NEYMAN, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *J. Roy. Statist*. *Soc*. *Ser. A* **109** 558–606.

PANEL ON PRIVACY AND CONFIDENTIALITY AS FACTORS IN SURVEY RESPONSE, COMMITTEE ON NATIONAL STATISTICS (1979). *Privacy and Confidentiality as Factors in Survey Response*. National Academy of Sciences, Washington, DC.

PLATEK, R., RAO, J. N. K., SMRNDAL, C. E. and SINGH, M. B. (1986). *Small Area Statistics*: *An International Symposium*. Wiley, New York.

REID, C. (1982). *Neyman—From Life*. Springer, New York. MR0680939

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

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York; Chapman & Hall, London. MR0116429

SCHUMAN, H., STEEH, C. and BOBO, L. (1985). *Racial Attitudes in America*: *Trends and Interpretations*. Harvard Univ. Press, Cambridge, MA.

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).

SENG, Y. P. (1951). Historical survey of the development of sampling theories and practice. *J. Roy. Statist. Soc. Ser. A* **114** 214–231.

SMITH, T. M. F. and SUGDEN, R. A. (1985). Inference and the ignorability of selection for experiments and surveys. *Bull. Int. Stat. Inst.* **51** 10.2-1–10.2-12. MR0886248

STEPHAN, F. F. (1948). History of the uses of modern sampling procedures. *J. Amer. Statist. Assoc.* **43** 12–39.

STOKES, S. L. (1986). Estimation of interviewer effects in complex surveys with application to random digit dialing. In *Proceedings of Second Annual Research Conference* 21–31. U.S. Bureau of the Census, Washington, DC.

TCHUPROV, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron* **2** 646–680.

TEPPING, B. J., HURWITZ, W. N. and DEMING, W. E. (1943). On the efficiency of deep stratification in block sampling. *J. Amer. Statist. Assoc.* **38** 93–100.

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. In *Proceedings of the Social Statistics Section* 227–237. Amer. Statist. Assoc., Washington, DC.

WILK, M. B. and KEMPTHORNE, O. (1955). Fixed, mixed, and random models. *J. Amer. Statist. Assoc.* **50** 1144–1167.

WILK, M. B. and KEMPTHORNE, O. (1956). Some aspects of the analysis of factorial experiments in a completely randomized design. *Ann. Math. Stat.* **27** 950–985. MR0087283

WOLTMAN, H. I., 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.

YATES, E. (1981). *Sampling Methods for Censuses and Surveys*, 4th ed. Macmillan, New York.

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, E. and COCHRAN, W. G. (1938). The analysis of groups of experiments. *J. Agric. Sci.* **28** 556–580.

ZARKOVICH, S. S. (1956). Note on the history of sampling methods in Russia. *J. Roy. Statist. Soc. Ser. A* **119** 336–338.

ZARKOVICH, S. S. (1962). A supplement to "Note on the history of sampling methods in Russia". *J. Roy. Statist. Soc. Ser. A* **125** 580–582.

DEPARTMENT OF SOCIOLOGY
STONY BROOK UNIVERSITY
STONY BROOK, NEW YORK 11794-4356
USA
E-MAIL: Judith.Tanur@stonybrook.edu