The Design of Pilot Studies: A Short Review

Lars Lyberg and Patricia Dean

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1989:22

Från trycket December 1989
Producent Statistiska centralbyrån, Utvecklingsavdelningen
Ansvarig utgivare Åke Lönnqvist
Förfrågningar Lars Lyberg, tel. 08-783 41 79

ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1989

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1989:22. The Design of pilot studies: A short review / Lars Lyberg, Patricia Dean
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1989-X101OP8922

The Design of Pilot Studies: A Short Review

Lars Lyberg and Patricia Dean

Statistics Sweden

Paper presented at the ASA Winter Conference in San Diego, CA
January 4-6, 1989

1. Introduction

Before beginning any discussion of pilot surveys or pilot studies, the term survey should be defined. This term can mean different things to different categories of researchers, even though they all could be avid producers and users of statistical information.

1.1 The notion of a survey

Survey literature fails to provide a single, precise definition of a survey. Dalenius (1974) approaches this problem by defining seven aspects that characterize a survey. These are summarized below.

1. A survey is performed on well-defined objects or sets of objects, referred to as a population.

2. The population has one or several measurable properties.

3. The population can be described by one or more parameters which are defined as functions of the measurable properties. To estimate these parameters, we observe a number (a sample) of the population's members.

4. To observe the population, we need a list of sampling units, i.e., a frame.

5. A sample of (ultimate) sampling units is selected from the frame in accordance with a sampling design. This sampling design specifies a probability mechanism that determines the sample size and the way in which the sample is selected. In special cases, the design may call for a total enumeration.

6. Observations are made on the sample in accordance with a measurement design, i.e., a measurement instrument and rules for its use.

7. The observations collected according to the above description are then used to compute estimates of the population parameters. These parameter estimates are used when making inferences about the population.

By this definition, any statistical study which can be characterized by these seven points is a survey.

1.2 The need for information when designing a survey

Designing a survey requires a great deal of prior information. When we say "prior" information, we mean that a survey designer must have a certain degree of knowledge about the population to be studied and even about the characteristics that are the topic of the statistical investigation. This leads to a paradox in which a survey researcher needs information that will not be available until the survey has been completed and, in some cases, the results tabulated. Obviously, this information must be obtained in other ways.

An efficient survey design calls for information on the variability of the population and data explaining this variability. One must also have some idea about the errors, costs, and administrative feasibility of the data collection mode and the data processing procedures. For example, the choice between alternative data collection modes and data processing procedures demands an extensive knowledge of the advantages and disadvantages of face to face vs telephone interviews, optical character recognition vs keypunching, manual vs automated coding, dependent vs independent verification, and the choice of one questionnaire over another. One must also be able to make the proper choice of sampling units (elements or clusters), sampling system (equal or unequal probabilities), and estimation system (the use or omission of auxiliary information). Without extensive information or knowledge of these methods and procedures, the choices facing the survey researcher become something of a gamble where the researcher hopes for the best possible outcome given his/her

expertise, experience, and the general survey conditions.

1.3 Information sources

The designer's experience and expertise are invaluable sources of information. Yet, to rely on experience only is insufficient. Relying only on experience usually leads to a monotonous approach when new surveys are to be designed. Design decisions tend to be handled in one way only, despite that every new design task presents different, even unique problems.

Similar surveys conducted earlier are also an invaluable source of information on variability, costs, administrative feasibility, and other general characteristics. However, the degree of similarity between variables is sometimes difficult to judge.

It is also beneficial for the survey researcher to review the literature that compares and evaluates different data collection modes and data processing methods. But nor should one rely too heavily on the information gained by scanning the literature; the astute reader will find a great deal of contradictory results. Some methodological studies are badly designed and confounding is not uncommon. Yet even with well-designed studies, the efficiency of a survey method depends on the specific study circumstances and usually all methods have advantages and disadvantages. The designer is always forced to evaluate how these factors affect his/her own survey design. This holds true even when the methods test has been conducted according to a strict experimental design.

Pretests and pilot surveys are another source of information and are the focus of this discussion. Pretests and pilot surveys are studies designed to generate information that can be useful when designing another survey, the main survey. Such studies are designed to generate specific information that is then used in planning.

Still another source of presurvey information is an evaluation study. An evaluation study is conducted simultaneously with or

directly after the main survey and the purpose of the evaluation is to produce some sort of quality statement about the main survey. There are two kinds of evaluation studies that are commonly conducted. In the one, various error components are estimated to provide the user with a data quality statement. In the other, the survey designer evaluates the various phases of the survey and this information is often used to improve a forthcoming survey. By conducting, for example, reinterviews, or by using preferred and more expensive survey procedures, it is possible to evaluate the various survey operations. Such studies are called producer-oriented evaluations.

In general, evaluation studies are strictly designed and provide estimates of the error components. The results of the evaluation study cannot, by definition, influence the main survey. The primary purpose of the evaluation study is to check the main survey. Sometimes, the results of an evaluation study can serve as an information bank when designing a similar survey in the future. It is, however, only in the most general sense that evaluation studies can be classified as a type of pilot study. Obviously, there is no need to use a method or technique that has proved untenable in an evaluation survey. On the one hand, pilot studies fall into the same general category with other producer-oriented evaluations. On the other hand, pilot surveys and evaluation studies are conducted in very different ways with very different purposes. Another characteristic that differentiates evaluation studies and pilot studies is that evaluation studies are expensive and not particularly common, whereas pre-survey pilot studies are common and need not be expensive.

2. Pilot Studies: A General Discussion

This review attempts to provide a discussion of pilot studies as a source of information when designing a survey. Pilot studies are sadly neglected in survey literature; textbooks on survey sampling deal only superficially with the topic. In some books, pilot studies are mentioned en passant while in others the matter is discussed in a few pages only. There is no obvious reason for

the lack of attention.

One explanation might be that pilot studies are seen as special cases of ordinary surveys and should be designed as such. In fact, one textbook author explicitly states that opinion. Nevertheless, the design of pilot studies deserves more attention than what it currently receives. The problems encountered in the design of pilot studies can be entirely different from those encountered in the design of regular surveys. The goal of a regular survey or a census is to provide sample estimates of parameters or census enumerations, whereas the pilot is likely to have many competing goals. The design of the pilot might be efficient for some of the goals of interest, but not for the others. Pilot studies have the option of using a random or a subjective sample, an additional complicating factor.

The same casual treatment that pilot study design has received in the literature is also seen in the pilots themselves. This is evident from even a cursive review of a number of pilot studies conducted over the past twenty years. The following rather basic flaws are found in an inordinate number of pilot studies. We find that inference is often replaced by intuition and that the goals, which are so important to the design, are loosely defined or even obscure. The pilot studies are often considered a mere warm-up for the main survey and not a valid survey research endeavor in itself. In these studies, cost efficiency is seldom an important feature. Pilot studies point out where problems might occur in the main survey, but no attempts are made to translate these problems into quantifiable terms or to establish their relative importance. Surprisingly often, the information gained from the pilot study is not used to make recommendations for the main survey; the magnitude of errors and costs often go unquantified. These and other shortcomings that occur in pilot studies would never be tolerated in regular surveys. Of course, there are many well-designed and well-implemented pilot studies, but the sheer number of badly designed ones should signal that something is amiss and deserves closer scrutiny.

2.1 Terminology

There is no generally accepted exclusive name for what we have referred to as pilot studies. Among the terms used by textbook and journal article authors, the following terms can be mentioned: pilot survey, pilot study, prepilot study, feasibility study, methodological study, dress rehearsal, pretest, formal test, informal test, experiments, and built-in experiments. Different authors mean different things with these labels. Although some differences in meaning do exist, the importance of these differences should not be exaggerated since in most cases the authors have not given much consideration to the inconsistencies in terminology. The following is a brief summary of various definitions and reasons for conducting pilot studies found in the literature and in personal communication with some researchers. We believe that the term pilot study is the most general and should be used to refer to the entire scope of studies conducted in preparation for another survey. Kalton (1982) suggests that the term pilot survey be reserved for pilot studies conducted on a larger scale, typically where the main surveys are government surveys that are going to be continuing. Pretests should be confined to testing of, for instance, a questionnaire on a small number of respondents, perhaps fewer than 50.

Other authors discuss the reasons for conducting pilot studies and propose situations that are appropriate for pilot studies. Kish (1965) advises that pilot studies be conducted when the designer is unfamiliar with the subject matter and that each problem area requires a separate pilot study. Financial constraints often result in pilot surveys that are too small. A pilot survey that is too small can produce equivocal and ungeneralizable results. In many cases, relying on experience and expertise produces better results than badly designed pilot studies. Zarkovich (1966) says that a pilot survey is a small scale survey designed to obtain the information needed to construct a rational survey design. The scale of the pilot should be proportional to the amount of information that is needed. For instance, the

scale of the pilot survey should be large if data on the subject matter are completely lacking. Cochran (1977) says that in undertaking a large-scale survey, particularly of unexplored material, it should be general praxis to conduct pretests and pilot and exploratory inquiries. The design of the pilot study should be tailored to the type of information needed. Bailar (1982) says that a pilot survey is relevant when the basic methodology of the main survey is already agreed upon and when information is needed on: the sample size required in the main survey, question feasibility, and variable costs of new procedures. Methodological studies usually explore the effects of alternative designs and generally require large samples.

Kasprzyk (1982) says that pilot surveys are sometimes confused with pretests. If new concepts, questions, or populations are to be included in a survey, in-depth pretesting of the survey instrument may be imperative to ensure that respondents are willing and able to provide the information requested. Pretesting should be used to refine the survey instrument on the major population subgroups (rather than estimating population parameters). Pilot surveys should be confined to testing the final versions of survey operations under operating conditions that resemble, as closely as possible, the environment of the main survey. If the pilot survey is large enough, it can include built-in experiments.

United Nations (1982) states that informal testing should always be a precursor to "voluntary" formal testing. Sometimes the main survey itself is considered a pilot survey and thus special pretests are not necessary.

Other opinions, comments, and discussions on these terms can be found in a number of textbooks, for example, Deming (1960), Yates (1960), Murthy (1967), Raj (1972), Konijn (1973), and Som (1973). Thus, it is difficult to arrive at a terminology that is self-evident. The following suggestion is based on a review of the literature and the personal communications mentioned and admittedly has its limitations.

Pretest and test	Usually a smaller study using informal qualitative techniques to explore the subject matter and the data collection instrument. Typically, a series of tests is needed to obtain the information required.
Pilot survey	Any survey that is designed and conducted to obtain information that can improve the main survey. It can be a single survey with multiple goals or a sequence of surveys. The design depends heavily on the survey's goals but will usually allow for reliable quantitative information and should be conducted at a time when the design of the main survey can still be changed.
Feasibility study	Formal or informal study of methods and procedures conducted when there are doubts about their practicability.
Embedded experiment, formal test, methodological study	An experiment, for instance, split-plot, can be made a part of a pilot survey or the main survey to test data collection modes, data or processing systems, and variations of a logical questionnaire. Such experiments should be strictly designed and usually require large sample sizes.
Dress rehearsal	A miniature of the main survey conducted close to the main survey to reveal weaknesses in the survey organization, to provide a base for improving survey instruments, and to provide realistic data for testing survey operations.
Main survey	The survey that is to be designed and conducted.
Evaluation	The results of the main survey are compared to those obtained by preferred procedures or reconciliation methods usually to produce a quality statement.

2.2 Information needs

The following is a list of planning problems that face the designer of a main survey. The list is not complete, but it covers most of the issues that lend themselves to pilot studies. Nevertheless, this list does show that there are many more areas in need of presurvey information than just question order, question wording, and nonresponse rates which are the most common areas of pilot

inquiry. Any survey designer should carefully try to identify his/her most important information needs before setting goals for a pilot study. Compiling a list of this kind and then ranking the needs can be a good base for a cost efficient pilot study design.

i. Data collection
- choice between different modes and combinations of modes
- problems due to type of survey, i.e., one-time, continuing, longitudinal
- timing, when to conduct the survey, length of recall period
- can the data actually be collected, sensitive topics, heavy respondent burden
- concepts and definitions
- use of special means, diary keeping, response cards, devices for randomized response
- confidentiality and ethics, informed consent, de-identification, information to respondents
- questionnaire development, layout, question wording, question order
- respondent rules
- special procedures, respondent screening operations performed by the interviewer, drop off - pick up procedures, address listing, use of new equipment including computers

ii. Sampling
- choice between total survey and sample survey
- construction and choice of sampling frames
- choice of sampling unit
- choice of estimation system
- causes of population variability
- size of variance, correlations, coefficients of variation, multistage variability, relative stratum variability, intraclass correlations
- magnitude of design effects
- sample size

iii. Field operations

- efficient use of field organization
- interviewer issues, performance, recruiting and training, instructions and manuals

iv. Data processing
- data capture: key punching, optical character recognition, personal computers
- coding: manual, automated, centralized and decentralized
- editing procedures: macro, micro, and extent of editing
- use of statistical quality control
- capacity

v. Nonsampling errors
- nonresponse rates: unit and item nonresponse
- methods for dealing with nonresponse: advance letters, incentives, adjustment
- other response errors: respondent, interviewer, data processing
- error variances

vi. Time and costs
- cost components
- cost per unit
- screening costs
- amount of time various activities demand

Of course, it is neither possible nor necessary to have information on all these aspects to produce a good design. There are also cases where financial, administrative, or technical constraints automatically reduce the number of alternatives. And, the choice of one design aspect might affect the number of options for other design aspects.

2.3 Typical pilot study methods

Questionnaire development is the most common area of pilot inquiry. Questionnaire development typically calls for a sequence of studies starting with informal tests and ending with more formal studies, for example, experiments. Questionnaire development usually

starts with a draft questionnaire which subject matter specialists and perhaps research colleagues attempt to complete. The rationale behind this procedure is that if people who are familiar with the subject matter or questionnaire development have problems understanding questions and instructions, then it is obvious that the general population will have even greater problems. Another measure to be taken in the early stages is to invite the interviewers to evaluate the questionnaire. Experienced interviewers need only read a questionnaire to detect weaknesses. Information from the interviewers can be obtained individually or via debriefing sessions where a group of interviewers discuss the questionnaire with the researcher. Such sessions are usually conducted after a few rounds of informal testing of the questionnaire. Similar activities can be performed with groups of respondents. Usually these informal early tests are conducted on a very small scale. In the U.S., it is not uncommon that the sample sizes of pilot studies conducted by government agencies are kept under 10 to avoid the need to obtain clearance from the Office of Management and Budget. Usually, however, at the early stages, 30-50 respondents are used to test a questionnaire.

After a number of refinements, more formal studies might be appropriate. Interpenetrating networks of subsamples is one example of a more formal study. It is also possible to use a formally designed experiment. Suppose we want to compare two alternative question wordings and that we have four interviewers at our disposal. We can then form blocks consisting of eight respondents. The respondents within each block should be as similar as possible according to certain background variables. The eight question-interviewer combinations are randomly assigned to the respondents in each block. This is a 2x4 randomized block, factorial design in which all differences can be tested simultaneously. It is also possible to obtain information on interaction between questions and interviewers. Variables used to discriminate between test alternatives might, for example, be item nonresponse, response distributions, or reported levels.

Brewer (1982) provides an example of census form development in

a developing country. In the 1966 Population Census of Papua-New Guinea four rounds of testing were conducted. The goal of the first round was to establish the feasibility of the entire operation, to identify the topics best suited for investigation, and get some indication of likely problems. For this study a single village, typical of the predominately rural population, was visited. The researchers learned that the questions they wished to ask were also of great interest to the villagers, such as, how old a certain individual was when another was born, or a woman's specific sequence of births and perinatal deaths. The information gained from the first round led to the use of probing rather than the minimum length forms which had been recommended by survey researchers with experience in other topics.

The second round was an elaborate sequence of pilot studies in the rural villages based on Graeco-Latin square experimental designs. These pilot studies sought to provide answers to, among others, the following questions. Could interviewers handle check-boxes, with or without skipping, and if not, how much time would be spent writing down the answers? How much training would the interviewers need before being able to handle the forms properly? What qualifications should the supervisors have and how many interviewers should they be responsible for? Should different forms be used in primitive areas, intermediate areas, and in areas of long standing European contact? This round was extremely useful. Since one of the intended questions caused a riot in one of the villages, the census would probably have collapsed had that question actually been asked.

The third round was conducted in urban and other nonrural areas and was mainly concerned with establishing feasibility and detecting specific problems. The appropriate length of the form was studied. The principal aim of the third round was to build confidence in the feasibility of the census operations for all areas of the country. The fourth round was quite small. It was used to check that the final census form was satisfactory.

When quantitative information is needed on, say, variability,

the pilot study's costs (time and money) can often outweigh its benefits. A pilot study estimating variability requires a large sample size and sometimes the size required is prohibitive. For instance, pilot studies are virtually useless for the estimation of intraclass correlations for multistage surveys. The sample sizes that are affordable are always far too small. If possible, some notion of the variability and similar information should be inferred from earlier surveys and the researcher's experience. A strongly held opinion is that several years of experience with different sample designs contributes more to the choice of optimal cluster sizes and subsampling fractions than any pilot study.

In the case where the researcher lacks data on important parameters, he/she can sometimes resort to inference based on empirical laws. The literature describes several variance functions based on cluster size. Proctor (1985) describes a method to fit H.F. Smith's empirical law to cluster variances for use in designing multistage sample surveys. Experience with intraclass correlation coefficients and design effects can also be helpful.

Cochran (1977) and some other textbooks discuss the estimation of population variances to determine sample sizes. A method that is not often used, because it is time-consuming, is to draw the sample in two steps. The first step is an srs of size n_1, leading to s^2, which is an estimate of the population variance S^2. Then s^2 can be used to obtain the required sample size n, provided n_1 is less than or equal to n. If a pilot survey using simple random sampling is used, the pilot can replace the first step described above. If the pilot survey sample is restricted in various ways, for instance, to clusters only, then we usually end up with an underestimate of S^2. Cochran also discusses a way of estimating the optimal subsample size in two-stage sampling by means of a pilot survey.

Yates (1960) discusses the case where it has already been decided that stratified sampling will be used in the main survey. If we cannot afford a simple random sample that is large enough to

allow variance calculations in each stratum, some form of multi-stage sampling can be used. In this sampling scheme, the first-stage sampling need not meticulously follow established survey praxis. Once the first-stage units are selected, the sampling is done in accordance with formal survey practice. If several stages are used, it is possible to cover selected areas with a density of the same order as that in the main survey. Then, various types and sizes of strata can be investigated. If multi-stage sampling is to be used in the main survey, the pilot study design becomes more complex. Since the variability of the first-stage units is so important, the pilot survey must include enough first-stage units and these units must also be selected at random from the first-stage strata. Such a complex pilot study design usually proves uneconomical.

2.4 Matters of inference

Pilot study designers often must use test results to make inference about a main survey to be conducted in the future on a different population and under different conditions. Such practices run contrary to a basic premise of inference which states that the results of a study are valid given the conditions under which the study was conducted. The fact that the pilot study is far more limited than the main survey leads to difficulties different from common survey problems and these difficulties are not dealt with in the survey literature. One exception is Brackstone (1976) who discusses the inferences drawn from the tests that preceded the 1981 Canadian Census of Population.

Informal testing can be used to indicate the areas, factors, or parts of a survey that could prove problematic for the main survey. Subjective information on, say, question wording, is obtained from interviewers, observers, and respondents. In this type of testing, an attempt is made to correct problems detected by one test before a new test is conducted. As the term suggests, informal tests are not designed to be evaluated with formal statistical methods. Informal tests in questionnaire development are almost always conducted with as few as 10 and as many as 300 respondents con-

centrated to a small number of sites. This type of informal testing does not permit inference to other sites and usually, inference is not relevant to the study. These informal procedures detect problems but do not allow for any evaluation of the magnitude of these problems. The choice of interviewers, respondents, and sites are not trivial choices. During the first tests in a sequence, it is preferable to use experienced interviewers because they are more apt at uncovering weaknesses and difficulties than less experienced interviewers. On the other hand, it is better to use a mix of interviewers with varying levels of experience in the later stages of informal testing. This mix of experienced and less experienced interviewers is closer to authentic survey conditions and is more likely to uncover problems that otherwise might have surfaced during the main survey.

It is a good idea to select "extreme," i.e., potentially problematic and unproblematic sites for the informal tests. For example, both inner city and rural areas, high and low income groups should be included in the pilot study. Testing the areas or groups that could produce unexpected results decreases the likelihood that surprises will pop up in the main survey. It is also important to get some indication of how well or how badly the main survey can proceed.

Informal testing is usually characterized by a lack of objective criteria for evaluating results. One common method is to study the response distribution and record the number of "don't knows" and the frequency of item nonresponse. Often the researcher is at a loss when it comes to interpreting what appears to be a problem. Poor question wording or question order might be the cause and only trial and error can resolve these problems. Informal testing should eventually lead to formal testing. Successive refinements should result in a number of alternatives that can be tested by, for example, split-plot experiments. Applications of informal testing are extensively dealt with in Nelson (1985). Informal testing can be applied to other fields as well. This is especially true for clerical operations such as coding.

Formal testing includes all studies that are designed to allow for estimates and comparisons. These studies can be separate or embedded experiments or carefully designed pilot surveys. It is absolutely essential that the test results are valid for the test environment. Otherwise inferences about a main survey environment become almost impossible.

Certainty about when inference is appropriate, i.e., whether the results of one study are applicable to another is a tricky problem. Brackstone (1976) discusses inference from three different perspectives, namely, time, universe, and discretionary conditions. Regarding the universe, it is seldom possible to use probability sampling of the entire population in a pilot study; in most cases costs would be prohibitive. On the other hand, with an affordable sample, the variances would probably be too large and in stratified sampling, sometimes even impossible to calculate. Even with pilot studies that are intended to be carefully designed, one might have to resort to some kind of subjective sampling or a mixture between subjective sampling and probability sampling.

One possible solution is to have a mixed sampling strategy, especially in multistage sampling. The primary sampling units can be chosen subjectively in a way that ensures some representativeness, that includes some expected problem areas, and that takes into account the practical constraints of sampling various primary units. The resulting inference is based partly on judgment and the incorporation of auxiliary data can prove helpful. Of course, sometimes it is possible to use probability sampling in pilot surveys that are conducted to estimate cost components, nonresponse rates, etc. The use of probability sampling circumvents many inference problems.

At the other extreme, inference about time related factors is based purely on judgment. This is especially true of the extensive census pilot studies that are so common in the U.S. and Canada. Most of these tests are carried out years before the actual day of the census. Of course, it must be extremely difficult to

foresee events that can make the conditions and circumstances of the main census significantly different from those of the test census. The only recourse is to incorporate expected changes into the judgmental inferences drawn from the test censuses.

The combination of the universe, the test schedule, the survey conditions, and inferential needs determine the test or pilot study that is possible. By and large, accurate estimates require random samples. For instance, the selection of interviewers for formal tests and other pilot studies should be a random sample of a pool of interviewers. In formal studies, interviewers should not be assigned on a voluntary basis or because their current workload is light enough to permit extra activities. One should not put too much faith into the results of pilot studies until the effects of nonrandom sampling, small sample sizes, limited number of sites or primary sampling units, seasonal variations, and number of alternatives tested are accounted for. The combined effect of large coefficients of variation, small sample sizes, and high nonresponse rates can be devastating for inference.

Brewer, Foreman, Mellor, and Trewin (1977) discuss the use of experimental designs in pilot surveys. If a pilot survey has several goals, the use of experiments can be efficient. Different options can be compared simultaneously. Usually, at least one combination works satisfactorily which can then reduce the number of tests.

For pilot studies, it is usually more sound to draw conclusions from estimation than from significance tests. A nonsignificant difference does not necessarily mean that the difference is unimportant and vice verse. Small sample sizes and false significances can explain some differences. As for experiments, Jabine (1981) recommends at least 500 objects and 10 agents per treatment as a rule of thumb.

2.5 Literature review

There are two kinds of literature on pilot studies. One treats the topic from a more general point of view, concentrating on design and methodology. This literature is scarce. Most textbooks on survey sampling mention pilot studies, but only superficially. Notable exceptions are Yates (1960), Kish (1965), Zarkovich (1966), Moser and Kalton (1972), and Cochran (1977). Nontextbook contributions are Brackstone (1976), Dunnell and Martin (1982), and Jabine (1981). Jabine provides the most comprehensive general discussion of the topic.

However, the literature on pretests and other pilot study activities for questionnaire development is vast and we can confidently conclude that the methodology for improving the survey instrument is well documented. Some important references for questionnaire development are Payne (1951), Sirken (1972), Lininger and Warwick (1975), Dillman (1978), Hoinville and Jowell (1978), Bradburn and Sudman (1979), Labaw (1980), Scherr (1980), Wright (1980), Schuman and Presser (1981), Hunt, Sparkman, and Wilcox (1982), Kalton and Schuman (1982), United Nations (1982), De Maio (1983), Jabine (1983), House (1985), Jabine (1985), Nelson (1985), Platek (1985). The use of experiments in pilot studies is discussed in Yates (1960), Jabine and Rothwell (1970), and Brewer et al. (1977).

At the other extreme, the literature covering pilot studies for survey operations other than questionnaire development is almost nonexistent. Of course, most textbooks on survey sampling mention these issues and they all have sections on sample size determination, but no individual topic is handled in depth. The scant literature that does describe pilot studies is found in internal documents from statistical agencies and conference proceedings. The topics discussed in the literature are found under the following headings: pilot surveys, experiments, feasibility studies, pretests, redesigns, methodology studies, and dress rehearsals. Some examples from this literature will be discussed in the next section.

3. Examples of Pilot Studies

As discussed earlier, all possible sources of information should be examined to obtain relevant and reliable information to plan efficient alternatives for a main survey. Sources like earlier surveys, special evaluation studies, methodological studies, feasibility studies, experience in general, expertise, and pilot studies in a broad sense are all important. These sources are intertwined and none can be discounted. For instance, when planning main surveys, the U.S. Bureau of the Census relies heavily on its expertise and experience with earlier surveys. This is especially true when estimating costs. Although numerous experiments and innovations are part of survey redesign and development, much of the required knowledge and skill is rooted in routine operating procedures. For example, when developing the Income and Survey Development Program (eventually resulting in the Survey of Income and Program Participation (SIPP)), the area probability sample was obtained from the current surveys' frame, the ultimate sampling unit being a cluster of addresses. The estimation procedure routinely includes a stage in which sample estimates are adjusted according to age, race, and sex population controls. In one development phase, intermittent approximations of variance estimates for the SIPP predecessor were obtained from the Current Population Survey March Supplement; adjustments were made to accommodate for differences between the two sample designs.

Evaluation studies are important, too. Every survey researcher and statistical agency would benefit from asking themselves what lessons had been learned from conducting specific surveys and how that knowledge can be used to improve future work. In this sense, evaluation studies can be classified as a special type of pilot study. Methodological studies that compare alternative procedures are frequent. When designed as experiments, these studies require large samples. If they can be embedded in an ongoing survey, cost efficiency is obtained. If adverse effects on the ongoing operations and results are expected, the study should be conducted separately. Embedded studies should be designed to permit quick termination if necessary.

In this section, we will provide examples of pilot studies and emphasize common design flaws, and of course we do not go into specifics when discussing flaws. We mention a few examples of pilot study work that is indeed excellent.

3.1 Well-designed pilot studies

In 1984, the U.S. Bureau of the Census and the U.S. National Center for Health Statistics conducted a feasibility study to investigate the use of random digit dialing techniques in the National Health Interview Survey. Nine specific goals were defined for this study including the estimation of costs and nonresponse rates, the evaluation of two alternative questionnaires to determine their effects on estimates, the identification of operational problems associated with administering the survey by telephone, and testing nonresponse and poststratification adjustments as well as different respondent rules. Ten different analysis projects were carried out and the major findings are found in U.S. Bureau of the Census and U.S. National Center for Health Statistics (1985).

Swensson and Tängdén (1979) describe a small embedded experiment. To investigate the effect of respondent burden on nonresponse rates, an experiment was conducted where individuals in one group were asked to participate in an extra survey that the other group was not asked to participate in. The extra respondent burden had a significant effect on the nonresponse rate and these results and other information eventually led to changes in the sampling for Statistics Sweden's continuing surveys of individuals and households. This study also resulted in a general policy that no one should be asked to participate in more than one survey every five years. The samples are coordinated in accordance with this policy.

The Income Survey Development Program was the forerunner to SIPP and exemplifies extensive pilot testing. Among the procedures and operations tested were: the use of administrative records as

sampling frames, a split-sample experiment with a short vs. a long form and a six-month vs. a three-month survey period, response errors, the forms' effect on interview length, estimation of nonresponse errors, estimation of response probabilities, comparisons of direct and indirect estimation of quarterly income, and field observation of interviewers. Descriptions of this extensive multigoal endeavor can be found in Coder (1980) and Kasprzyk (1988).

Catlin and Ingram (1979) discuss Canadian victimization surveys that are based on reported crimes; they maintain that these surveys underestimate the number of victims. It seems natural to try to complement the register-based data collection with different surveys. A sample of victims was interviewed face to face or by telephone. The design is interesting and includes experiments, reverse record checks, informal procedures, and the use of an extra sample of the population included to keep interviewers unaware of whether the respondent came from the register of victims or not. The results from one pilot study led to the decision to conduct another study. The studies showed that telephone surveys can be used to collect reliable information on victimization.

Good pilot study practices are also seen in the following work. Bailey, Moore, and Bailar (1978) study interviewer variability in an embedded experiment. Fernandez and McKenney (1980) developed a question aiming at identifying the Hispanic population in the 1980 U.S. Census of Population and Housing. Lyberg (1978) used a formal experiment to investigate the feasibility of simplifying the diary keeping in the 1978 Swedish Household Expenditure Survey. Bushery (1981) evaluated the recall biases for different reference periods in the U.S. National Crime Survey. Brewer and Whittington (1969) describe the exploratory and experimental activities that were used to refine the 1966 Sample Population Census of the Territory of Papua and New Guinea.

3.2 Some common design flaws

When reviewing pilot study reports, the number of design flaws is indeed amazing. It is common that the goals of the study are not specified in detail and that the final report lacks tangible recommendations. Procedures are inadequately described and some studies are designed so that confounding is inevitable. In these cases, one can only conclude that these studies have not been conducted according to standard survey methodology. Financial and time constraints and the attitude that pilot studies are for in-house-use-only can lead to pilot studies that are conducted more as a warm-up rather than a cost efficient procedure.

We also found flaws in pilot studies that were highly subtle, and in this way more insidious than the more blatant flaws. Here we discuss a few examples of subtle yet serious flaws and reiterate some of our more important points.

The choice of study areas or sites is not always well-founded. We have seen that subjective samples of primary sampling units can be both necessary and often the best choice. On the other hand, choosing areas reputed to be extreme problem or nonproblem areas implies that the choice is based on knowledge of the issues rather than administrative convenience. To base such a choice on anything other than knowledge and reasonable expectations is to make a great error.

Sometimes excessive coding, editing, and tabulation are done without a specified purpose.

A researcher can end up with a lot of tables containing results generated by nonrandom samples. Sometimes, the sampling strategy per se is flawed. But the more serious error is the researcher's tendency to discuss the results from a subject-matter point of view rather than from a methodological point of view.

Sometimes it seems as if the sample size of the pilot study has a magic all its own. It is claimed that a certain number of cases

"must" be reached to make a proper analysis possible. This might not always be wrong, but an error is committed when there is no rationale for conducting successive sampling until a certain sample size is obtained.

Sometimes a small pilot study reflects low ambitions rather than cost efficiency. If a new data collection mode is in question, it is not enough to study, say, just the questionnaire and the nonresponse rate. Sometimes the number of interviewers is too small. We have come across one study where just a single interviewer participated. Nonexperimental studies where the number of interviewers is smaller than ten are not uncommon. We have also come across studies involving just one hospital, one industrial plant, one coder, etc. What is most upsetting about these studies is that the goals have been set far above the preliminary exploratory level which sometimes could justify these minimal sizes.

If the general methodology is agreed upon (a decision that should be based on extensive information) it is valuable to carry out a dress rehearsal to refine the procedures. But surprisingly often, only one alternative has been considered prior to the dress rehearsal, even though the available information is vague. We are left without knowing whether better or worse alternatives exist.

One study indicated several problems with population definition. The study was correctly designed but those responsible thought that the problems could be handled in the main survey. The eventual main survey suffered from exactly the same problems as indicated by the pilot study plus a number of new problems that the pilot study did not cover. The results could not be published. What we learn from this is that sometimes the proper decision is not to conduct the main survey.

While going through approximately a hundred pilot study reports we have come across deficiencies that we think will never be repeated. There is no benefit to be reaped from commenting on these.

4. Acknowledgements

In 1982, Barbara Bailar, Tore Dalenius, Ken Brewer, Graham Kalton, Bob Groves, Dan Kasprzyk, Gordon Brackstone, and Dick Platek kindly devoted time to discussing pilot study issues with Lars Lyberg. This paper benefited greatly from their comments and ideas and we would like to thank them for their help.

5. References

Bailar,B.(1982): Personal communication.

Bailey, L., Moore, T.F., and Bailar, B.(1978): An Interviewer Variance Study for the Eight Impact Cities of the National Crime Survey Cities Sample. Journal of the American Statistical Association, Vol. 73, pp 16-23.

Brackstone, G.J.(1976): Drawing Inferences From Test Results. Memo, Statistics Canada.

Bradburn, N.M. and Sudman, S.(1979): Improving Interview Method and Questionnaire Design. Jossey-Bass Publishers, San Francisco.

Brewer, K.R.W. and Whittington, A.J.(1969): The 1966 Sample Population Census of the Territory of Papua and New Guinea. Review of the International Statistical Institute, Vol.37:2, pp 149-164.

Brewer, K.R.W.(1982): Personal communication.

Brewer, K.R.W., Foreman, E.K., Mellor, R.W., and Trewin, D.J.(1977): Use of Experimental Design and Population Modelling in Survey Sampling. Bulletin of the International Statistical Institute, pp 1-18.

Bushery, J.M.(1981): Recall Biases for Different Reference Periods in the National Crime Survey. Memo, US Bureau of the Census.

Catlin, G. and Ingram, S.(1979): Report on Canadian Victimization Survey Methodological Pretests. Memo, Statistics Canada.

Cochran, W.G.(1977): Sampling Techniques. Wiley, New York.

Coder, J.F.(1980): Some Results From the 1979 Income Survey Development Program Research Panel. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp 540-545.

Dalenius, T.(1974): Ends and Means of Total Survey Design. Forskningsprojektet FEL I UNDERSÖKNINGAR, Department of Statistics, Stockholm University.

De Maio, T.J.(Ed.)(1983): Approaches to Developing Questionnaires. Statistical Policy Working Paper 10, Office of Information and Regulatory Affairs, Office of Management and Budget.

Deming, W.E.(1960): Sample Design in Business Research. Wiley, New York.

Dillman, D.A.(1978): Mail and Telephone Surveys: The Total Design Method. Wiley, New York.

Dunnell, K. and Martin, J.(1982): Piloting: Purposes and Evaluation. Survey Methodology Bulletin, No.14, pp. 36-41.

Fernandez, E.W. and McKenney, N.R.(1980): Identification of the Hispanic Population: A Review of Census Bureau Experiences. Memo, US Bureau of the Census.

Hoinville, G., Jowell, R., and Associates(1978): Survey Research Practice. Heinemann Educational Books, London.

House, C.C.(1985): Questionnaire Design with Computer Assisted Telephone Interviewing. Journal of Official Statistics, Vol.1, No.2, pp 209-219.

Hunt, S.D., Sparkman, R.D., and Wilcox, J.B.(1982): The Pretest in Survey Research: Issues and Preliminary Findings. Journal of Marketing Research, Vol.XIX, pp 269-273.

Jabine, T.B.(1981): Guidelines and Recommendations for Experimental and Pilot Survey Activities in Connection With the Inter-American Household Survey Program. Paper prepared for The Inter-American Statistical Institute.

Jabine, T.B.(1985): Flow Charts - A Tool for Developing and Understanding Survey Questionnaires. Journal of Official Statistics, Vol.1, No.2, pp 189-207.

Jabine, T.B. and Rothwell, N.D.(1970): Split-Panel Tests of Census and Survey Questionnaires. American Statistical Association, Proceedings of the Social Statistics Section, pp 4-13.

Jabine, T.B.(1983): Teaching Questionnaire Design. Statistical Review, No.5, pp 157-166.

Kalton, G.(1982): Personal communication.

Kalton, G. and Schuman, H.(1982): The Effect of the Question on Survey Responses: A Review. Journal of the Royal Statistical Society, Series A, 145, Part I, pp 42-73.

Kasprzyk, D.(1988): The Survey of Income and Program Participation: An Overview and Discussion of Research Issues.
R & D Report 1988:14, Statistics Sweden.

Kasprzyk, D.(1982): Personal communication.

Kish, L.(1965): Survey Sampling. Wiley, New York.

Konijn, H.S.(1973): Statistical Theory of Sample Survey Design and Analysis. North-Holland.

Labaw, P.(1980): Advanced Questionnaire Design. Abt Books, Cambridge, Mass.

Lyberg, I.(1978): Kombination av bokföringsformer i HBU 78. Statistiska metodproblem i hushållsundersökningar, No.XII, Statistics Sweden (In Swedish).

Moser, C. and Kalton, G. (1972): Survey Methods in Social Investigation. Basic Books, New York.

Murthy, M.N.(1967): Sampling Theory and Methods. Statistical Publishing Society, Calcutta.

Nelson, D.D.(1985): Informal Testing as a Means of Questionnaire Development. Journal of Official Statistics, Vol.1, No.2, pp 179-188.

Payne, S. (1951): The Art of Asking Questions. Princeton University Press.

Platek, R.(1985): Some Important Issues in Questionnaire Development. Journal of Official Statistics, Vol.1, No.2, pp 119-136.

Proctor, C.H.(1985): Fitting H.F. Smith's Empirical Law to Cluster Variances for Use in Designing Multi-Stage Surveys. Journal of the American Statistical Association, Vol.80, No. 390, pp 294-300.

Raj, D.(1972): The Design of Sample Surveys. McGraw-Hill, New York.

Scherr, M.G.(1980): The Use of Focus Group Interviews to Improve the Design of an Administrative Form: A Case Study at the Social Security Administration. Memo, US Social Security Administration.

Schuman, H. and Presser, S.(1981): Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context. Academic Press, New York.

Sirken, M.G.(1972): Designing Forms for Demographic Surveys. Laboratories for Population Statistics, Manual Series No.3, The University of North Carolina.

Som, R.K.(1973): A Manual of Sampling Techniques. Heinemann Educational Books, London.

Swensson, B. and Tängden, E.(1979): Uppgiftslämnarbördans effekt på bortfallsstorleken- en empirisk studie. Memo, Statistics Sweden (In Swedish).

United Nations(1981): Handbook of Household Surveys. Introduction and Part One. General Survey Planning and Operations. Statistical Office, United Nations.

United Nations(1982): National Household Survey Capability Programme. Statistical Office, United Nations.

US Bureau of the Census and the US National Center for Health Statistics(1985): The Results of the 1984 NHIS/RDD Feasibility Study: Final Report. Paper submitted to the NCHS-Census Joint Steering Committee on Telephone Surveys.

Warwick, D.P. and Lininger, C.A.(1975): The Sample Survey. Theory and Practice. McGraw-Hill, New York.

Wright, P.(1980): Strategy and Tactics in the Design of Forms. Visible Language, XIV2, pp 151-193.

Yates, F.(1960): Sampling Methods for Censuses and Surveys. Griffin, London.

Zarkovich, S.S.(1966): Quality of Statistical Data. FAO, Rome.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports, Statistics Sweden, are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown (beige) covers).

Reports published earlier during 1989 are:

1989:1 (grön) **Går det att mäta produktivitetsutvecklingen för SCB? (Rune Sandström)**

1989:2 (grön) Slutrapporter från U-avdelningens översyn av HINK och KPI **(flera författare)**

1989:3 (grön) A Cohort Model for Analyzing and Projecting Fertility by Birth Order **(Sten Martinelle)**

1989:4 (beige) **Abstracts I** - Sammanfattningar av metodrapporter från SCB

1989:5 (gul) On the use of Semantic Models for specifying Information Needs **(Erik Malmborg)**

1989:6 (grön) On Testing for Symmetry in Business Cycles **(Anders Westlund and Sven Öhlén)**

1989:7 (grön) Design and quality of the Swedish Family Expenditure Survey **(Håkan L Lindström, Hans Lindkvist and Hans Näsholm)**

1989:8 (grön) Om utnyttjande av urvalsdesignen vid regressionsanalys av surveydata **(Lennart Nordberg)**

1989:9 (grön) Variations in the Age-Pattern of Fertility in Sweden Around 1986 **(Michael Hartmann)**

1989:10 (grön) An Application of Generalized Precision Functions in the 1985 Swedish Family Expenditure Survey **(Håkan L Lindström and Peter Lundquist)**

1989:11 (gul) Statistics production in the 90's - decentralization without chaos **(Bo Sundgren)**

1989:12 (grön) Översyn av urvalen i de objektiva skördeuppskattningarna **(Erling Andersson)**

1989:13 (gul) ADBs roll i statistiken **(Bo Sundgren)**

1989:14 (grön) Krisgruppsarbetet och räknarexperimentet i HUT 88 **(Håkan L. Lindström)**

1989:15 (grön) On evaluation of surveys with samples from the revised Zimbabwe master sample frame **(Bengt Rosén)**

Var god vänd!

1989:16 (grön) Bortfallsbarometern nr 4 **(Sonia Ekman, Tomas Garås, Hans Pettersson, Monica Rennermalm)**

1989:17 (beige) **Abstracts II** - sammanfattning av metodrapporter från SCB

1989:18 (gul) Conceptual modelling as an instrument for formal specification of statistical information systems **(Bo Sundgren)**

1989:19 (grön) Rapport från ett besök vid U.S. Bureau of the Census **(Gösta Forsman)**

1989:20 (gul) Regler och rekommendationer vid konstruktion av användargränssnitt **(Hans Irebäck)**

1989:21 (beige) **Abstracts III** - sammanfattning av metodrapporter från SCB

Kvarvarande **beige** och **gröna** exemplar av ovanstående promemorior kan rekvireras från Elisabet Klingberg, U/STM, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 78.

Kvarvarande **gula** exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.

CATARINA ELFFORS

A/BF-S