AN INVESTIGATION OF RESPONSE VARIANCE IN SAMPLE SURVEYS

Colm Aongus O'Muircheartaigh

Department of Statistics
London School of Economics and Political Science

Submitted for examination in October of 1999 for the degree

Doctor of Philosophy

1999

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

The dissertation considers response variance in sample surveys in the broader context of survey quality and survey error. Following a historical review of the evolution of both the terms and the concepts a brief overview is given of earlier research in the area. The principal content of the dissertation draws on investigations carried out by the author over the last thirty years.

There are three separate strands of argument, each associated with a particular approach to the analysis. First there is the descriptive (simple diagnostic) orientation of establishing the circumstances under which (or if) response variance arises, the associated issue of how it should be accommodated in analysis - primarily estimating the impact on the variance of univariate statistics - and an assessment of its likely order of magnitude. Second, there is the model-assisted orientation which attempts to decompose the effects into their constituent parts: one approach is to incorporate the correlating source (cluster or interviewer for example) as a term or terms in other models that we are estimating so that the effect is incorporated into the estimation of these models; the other is to model the response error itself -- in doing this we are trying to decompose it into its constituent parts. Third, and most radical, is to view error as information. By conceptualizing the process that generated the errors as a substantive process rather than as a set of nuisance effects we can extract from the results of the process information about both the process and the subject matter. Any particular piece of analysis may include any combination of these three approaches.

The dissertation draws on special studies incorporated into a number of major sample surveys. Two principal data sets are involved. The first arises from a special investigation of response error carried out in conjunction with the World Fertility Survey; the second is the reinterview data set from the Current Population Survey carried out by the US Bureau of the Census. Four other surveys are used; an absenteeism survey in Ireland, two cross-sectional British surveys (one on Noise Annoyance, the other on Physical Handicap), and a British panel survey (the British Household Panel Survey).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 THE HISTORICAL ORIGINS OF QUALITY ASSESSMENT IN SURVEYS

This chapter traces the evolution of the concept of error in social surveys and places total variance as defined by social statisticians in that broader context[1].

1.1 Origins of Survey Assessment

In considering the different contexts in which early users of surveys operated, and the different perspectives they brought to their operations, it is difficult to find common criteria against which to measure the success of their endeavours. The history of surveys (in their modern sense) goes back only 100 years, but from the outset there was a great diversity in the settings, topics, philosophies, and executing agencies involved. In the initial stages there was no particular distinction drawn between the issues of survey design and execution and the issues of error in surveys.

The concept of quality, and indeed the concept of error, can only be defined satisfactorily in the same context as that in which the work is conducted. To the extent that the context varies, and the objectives vary, the meaning of error will also vary. I propose that as a definition of error we adopt the following: *work purporting to do what it does not do*. Rather than specify an arbitrary (pseudo-objective) criterion, this redefines the problem in terms of the aims and frame of reference of the researcher. It immediately removes the need to consider *true value* concepts in any absolute sense, and forces a consideration of the needs for which the data are being collected. Broadly speaking, every survey operation has an objective, an outcome, and a description of that outcome. Errors (quality failures) will be found in the mismatches among these elements.

There are three distinct strands in the historical development of survey research: governmental/official statistics; academic/social research; and

---

1

A somewhat longer version of some of the material in this chapter can be found in O'Muircheartaigh (1997); some similar issues are discussed in O'Muircheartaigh (1999).

commercial/advertising/market research. Each of these brought with it to the survey its own intellectual baggage, its own disciplinary perspective, and its own criteria for evaluating success and failure.

The International Statistical Institute was the locus of debate for official statisticians at the end of the 19th century when Kiaer, director of the Norwegian Bureau of Statistics, presented a report of his experience with "representative investigations" and advocated further investigation of the field. In this context the evaluation of surveys was largely statistical and the survey was seen as a substitute for complete enumeration of the population. Bowley - the first professor of statistics at the London School of Economics and Political Science - through his work on sampling (1906 and 1926) and on measurement (1915) was one of the principal figures in the development of the scientific sample survey. This became and has remained the dominant methodology in the collection of data for government, and the government sample survey agency became an important purveyor of data both to politicians and to statesmen. Symptomatic of their genesis, these agencies tended to be located in the national statistical office, and their professional staff tended to be trained in mathematics or in statistics. Here the concept of error became synonymous with the variance of the estimator (essentially the variance of the sampling distribution following Neyman's influential paper in 1934 (Neyman, 1934)). This equivalence of quality and variance and its measurement by repeated sampling, with some acknowledgement of bias, was confirmed by the work of Mahalanobis in India, reported in the mid-1940s (see Mahalanobis, 1944, 1946), and in particular by his design of interpenetrated samples for the estimation of fluctuations or variability introduced by fieldworkers and others. The influence of statisticians on the conceptualization of error and its measurement has continued in this tradition, and can be found in all the classic texts of survey sampling (Yates, 1949; Cochran, 1953; Hansen, Hurwitz and Madow, 1953; Kish, 1965). In this tradition the term "error" has more than one meaning (see for example Groves (1989) but it is used loosely to describe any source of variation in the results or output or estimates from a survey.

While recognising the powerful position occupied by the scientific sample survey in social research, it is worth noting that Kiaer's proposal to the ISI in 1895 was not universally

welcomed, and would almost certainly have been rejected had it not been for the support of the monographers whose work consisted of the detailed examination of one or a small number of cases (what might today be called the *case study* approach).

The involvement of the monographers in the debate at the ISI is interesting particularly because it provides a link to the second major strand in the development of surveys. This was the *Social Policy* and *Social Research* movements, whose beginnings are perhaps best represented by Booth's study, from 1889 to 1903, of poverty in London, and the Hull House papers in the USA in 1892. Though not in any way a formal or organised movement, there were certain commonalities of approach and objectives across a wide range of activities. The goal of this movement was social reform, and the mechanism was community description. Here the success or failure of the activity was the effect the findings had on decision makers and politicians.

The principal influences on this group were the social reform movement and the emerging sociology discipline. Some of the pioneers of sample surveys spanned both official statistics and the social survey; in particular, Bowley (who made a substantial contribution to the development of survey sampling) produced a seminal work on social measurement in 1915 which helped define the parameters of data quality and error for factual or behavioral information. Bogardus (1925), Thurstone (1928), and Likert (1932) provided scientific approaches to the measurement of attitudes. In this field the disciplinary orientation was that of sociology and social psychology, with some influence from social statistics and psychometrics. Likert, who was subsequently the founding director of the Institute for Social Research at the University of Michigan in 1946, reflected the same practical orientation as the early pioneers in the Social Research movement in his later work on organizations (though with extensive use of attitude measurement).

The third strand arose from the expansion of means of communication and growth in the marketplace. From modest beginnings in the 1890s (Gale and others), there was a steady increase in the extent of advertising and a development and formalization of its companion, market research. The emphasis was on commercial information, originally in the hands of

producers of goods and services and collected and evaluated informally (Parlin, 1915); market research, however, developed gradually into a specialized activity in its own right.

Here the effect of psychologists was particularly strong. The work of Link and others in the Psychological Corporation was influential in providing an apparently scientific basis for measurement in the market research area. For those psychologists, experimental psychology took precedence over social psychology. The terminology and the approach were redolent of science and technology. The term "error" was not used explicitly; rather there was a description of *reliability* and *validity* of instruments. This contrasts particularly with the "error" orientation of the statisticians.

Thus the field of survey research as it became established in the 1940s and 1950s involved three different sectors - government, the academic community, and business; it had three different disciplinary bases - statistics, sociology and experimental psychology; and it had developed different frameworks and terminologies in each of these areas.

1.2 Framework

In general in describing data quality or errors in surveys, models concentrate on the survey operation itself, in particular on the data collection operation. The models may be either mathematical (presenting departures from the ideal as disturbance terms in an algebraic equation) or schematic (conceptual models describing the operational components of the data collection process). The conceptual models focus on the interview as the core of the process. Building on the work of Hyman (1954), Kahn and Cannell (1957), Scheuch (1967) and others, Sudman and Bradburn present one of the more useful of these in their book on response effects in surveys (Sudman and Bradburn, 1974). This (schematic) model (a simplified version of which is presented in figure 1.1) presents the relationship among the interviewer, the respondents and the task in determining the outcome of the survey interview. The elaborated model identifies the potential contribution of a number of the key elements in each of these to the overall quality of the survey response.

Figure 1.1 about here

- The *interviewer*, as the agent of the researcher, is seen to carry the lion's share of responsibility for the outcome of the data collection process. Sudman and Bradburn distinguish three elements: the formal constraints placed on the interviewer; the actual behaviour of the interviewer; and the extra-role characteristics of the interviewer.

- The *respondent* has not generally been examined as a source of error (apart from a general complaint about poor performance of his/her task). Survey research has tended to take the respondent for granted, though many of the early writers referred to the need to motivate the respondent. The overall approach has, however, been to consider the respondent an obstacle to be overcome rather than an active participant in the process.

- In general, models of response errors in surveys focus on the *task*, which is constrained and structured to accomplish the research goals - in particular to provide the data necessary for analysis. The task includes the length and location of the interview, the question wording, questionnaire construction, the types of data sought, and their implications in terms of memory and social desirability.

- The *responses* are the outcome of the data collection exercise, and the raw material for survey analysis. Most survey analyses treat these as free from error; the statistical approach to measurement error considers the response to be a combination of the *true value* of the data for the individual plus a disturbance described as a *response deviation* or *response effect.*

It is clear that any model of the survey process, and therefore any general model of survey error, will have to include these elements. It is not, however, sufficient to consider these elements, as they do not take into account the context of a survey nor can they distinguish among different survey objectives. To compare the different approaches to survey research

described in section 1.1, however, it is necessary to provide an over-arching framework that encompasses the concerns of all three major sectors.

One possible framework draws on some ideas presented by Kish in his book on statistical design for research (Kish, 1987). Kish suggests that there are three issues in relation to which a researcher needs to locate a research design; I propose that a similar typology could be used to classify the dimensions that would encompass most sources of error. Each of these "dimensions" is itself multi-dimensional; they are *representation*, *randomization*, and *realism*. The model is represented in figure 1.2.

Figure 1.2 about here

As survey research deals with applied social science our understanding of measurement in surveys must also be grounded in actual measures on population elements. Social theory does not have this requirement, nor indeed does statistical theory. At this empirical level, however, the strength and even the direction of relationships between variables are always conditional on the elements, and thus it is critical that any conclusions from a survey should be based on a suitable set of elements from the population, and that comparisons between subclasses of elements should be based on comparable subsets of elements. *Representation* involves issues such as the use of probability sampling, stratification, the avoidance of selection bias, and a consideration of nonresponse. In general we do not believe that any finding in social science will apply uniformly to all situations, in all circumstances, for all elements of the population. Indeed a good deal of social science is dedicated to understanding the ways in which differences occur across subgroups of populations or between populations. *Representation* reflects this set of concerns with regard to the elements included in the investigation. In particular it refers to the extent to which the target population is adequately mirrored in the sample of elements. In a perfectly specified model, there would be no need to be concerned about which elements from the population appeared in the sample. In the absence of complete and perfect specification of a model, (with all variables with potential to influence the variables or relationship under consideration being included), the notion of representation specifically covers the appropriate representation of

domains (or subclasses), the avoidance of selection bias, and the minimization of differential nonresponse. The term representative sampling has a chequered history in statistics (see, for instance, Kruskal and Mosteller, 1980; O'Muircheartaigh and Soon, 1981). It carries with it an aura of general (possibly) unjustified respectability; it can be taken to mean the absence of selective forces (that could lead to selection biases); its original connotation was that of a miniature or mirror of the population; it has sometimes been seen as a typical or ideal case; it can imply adequate coverage of the population (cf. stratification); its highest manifestation is in probability sampling, which is the approach in academic and (most) governmental research.

Randomization (and its converse in this context, *control*) covers issues of experimentation and control of confounding variables. Though surveys rarely involve the use of formal experiments for substantive purposes, the identification of sources of measurement error (distortion in the data) and the estimation of the magnitudes of these "errors" frequently do. Randomization is used to avoid, or reduce the probability of, spurious correlations or mis-identification of effects. (It may be worth pointing out that randomisation (or at least random selection) is also used to achieve representation in probability sampling.)

Realism arises as an issue in this context in two ways. *Realism in variables* concerns the extent to which the measured or manifest variables relate to the constructs they are meant to describe; *realism in environment* concerns the degree to which the setting of the data collection or experiment is similar to the real-life context with which the researcher is concerned. The survey context may be contrasted with observational studies in which both the variables and the environment are closer to the reality we would like to measure. These dimensions are related to the ideas of *internal validity* and *external validity* used by Campbell and Stanley (1963) and others in describing the evaluation of social research. The validity of a comparison within the context of a particular survey is the realm of internal validity; the extent to which an internally valid conclusion can be generalized outside that particular context is the realm of external validity.

In the following sections the different components of the response process are presented.

Each of them concentrates on a different element of the basic model. Section 1.3 presents the perspective of official (government) statistics and concentrates on the *responses*; it is this tradition still followed by the *hard science* school of survey research. Section 1.4 considers the elements of the *task*; section 1.5 takes as its focus first the *interviewer*, then the *respondent* and the interrelationship between them. In sections 1.4 and 1.5 most of the contributions to progress have been made by either the psychologists involved in market and opinion research, or by the sociologists and social psychologists involved in social and policy research. In section 1.6 some recent developments are used to illustrate how measurement error in surveys is being reconsidered. Section 1.7 presents the plan of the dissertation.

1.3 The Effect of the Sampling Perspective - Statistical Analysis of the Responses: Official Statistics and Survey Statistics

The sample survey was seen by Kiaer and its other originators in government work as an alternative to complete enumeration necessitated by the demand for more detail and more careful measurement. In 1897 Kiaer wrote *"In order to arrive at a deeper understanding of the social phenomena ... it is necessary to ... formulate a whole series of special questions .... prohibitive to conduct a complete survey of the population of a country, indeed even one for all the inhabitants of a large town"* (p. 38). It was the necessity to *sample* that brought about the difference between the survey and the usual government enquiry, and it was the errors that might contaminate the results because of this that became the focus of concern for the first generation of statisticians and others involved with government surveys. Kiaer suggested *replication* - simply constructing a set of comparable subsamples (in essence repeating the sampling operation) - as the means of evaluating the survey results (p. 51); this was, as far as I know, the first *total variance model.*

This approach was taken on board by Bowley and other statisticians and culminated in the classic 1934 paper by Neyman to the Royal Statistical Society "On the Two Different Aspects of the Representative Method" which crystallised the ideas in his concept of the *sampling distribution* - the set of all possible outcomes of the sample design and sampling

operation. The quality of a sample design, and thus a sample survey, could be encapsulated in the *sampling error* of the estimates; it is worth noting that though the general term "error" was used, this was a measure purely of variance or variability, and not necessarily a measure of error in any general sense. In particular, *bias* (or systematic error) was not a primary consideration, except as a technical issue in the choice among statistical estimators.

The Kiaer-Bowley-Neyman approach produced the sequence of texts on sampling which have defined the social survey field for statisticians ever since. The sequence of classic sampling texts began with Yates (1949, prepared at the request of the United Nations Sub-Commission on Statistical Sampling for a manual to assist in the execution of the projected 1950 World Census of Population), and Deming (1950), followed by Hansen, Hurwitz and Madow (1953), Cochran (1953), and Sukhatme (1953), and concluded with Kish (1965). With these texts survey statisticians defined their field as that of measuring the effect of sample design on the imprecision of survey estimates. Where other considerations were included they tended to be relegated to a subsidiary role, or confined to chapters towards the end of the book. The texts vary a good deal in terms of the relative weight given to mathematical statistics; the most successful as a textbook, however, Cochran (2nd edition 1963, 3rd edition 1977) was the most mathematical and the least influenced by nonsampling and nonstatistical concerns.

A second strand was present in the work of Mahalanobis in India. He, like Kiaer, advocated the use of replication, using what he called *interpenetrating samples*, to estimate the precision of estimates derived from a survey. He defined these as "independent replicated networks of sampling units". He was, moreover, the first statistician to emphasise the *human agency* in surveys (1946, p. 329); he classified errors as those of sampling, recording, and physical fluctuations (instability). To estimate variance, he advocated that different replicates should be dealt with by "different parties of field investigators" so that human error as well as sampling errors would be included in the estimates of precision; he also carried out tests of significance among replicates. Mahalanobis may also be credited with perceiving that an additional advantage of partial investigations (using his interpenetrating samples) was that they facilitated the estimation of error, something previously not a part of the reports of

government agencies. Indeed one of his early evaluations, based on sampling by the Indian Civil Service between 1923 and 1925, showed a bias in the estimation of crop yields (1946, p. 337).

Replication remains the primary instrument of statisticians when dealing with error. The traditional division between *variance* - the variability across replications, however defined - and *bias* - systematic deviation from some correct or true value - still informs the statistician's approach to error. Replication (or Mahalanobis's interpenetration) is the method of producing *measurability* in the sense of being able to estimate the variability of an estimate from within the process itself. In sampling this was brought about by selecting a number of sampling units independently and using the differences among them as a guide to the inherent stability or instability of the overall estimates. For simple response variance, the statisticians simply repeated the observations on a subset of respondents and compared the outcomes; this is usually called a reinterview programme; this gives replication in the sense of repetition. In the context of interviewer effect, the replication is within the survey operation and is brought about by constructing comparable subsets of cases and comparing them. To measure interviewer effect, respondents are allocated at random to different interviewers and the responses obtained are compared. Statistical theory tells us how much variability we could expect among these interviewer workloads if there is no systematic interviewer effect on the responses. To the extent that the variation is larger than that predicted by the null model, we attribute the effect to the interviewers.

The early 1960s saw the next step forward in statisticians' consideration of survey error. Hansen, Hurwitz and Bershad (1961) in a seminal paper presented what became known as the "U.S. Census Bureau model" of survey error. They defined the *essential survey conditions* as the stable characteristics of the survey process and the survey organisation carrying it out; variance was defined relative to those essential survey conditions. The observation is seen as being composed of two parts, its *true value*, and a deviation from that value - the *response deviation*. Though Hansen and his colleagues were well aware that the notion of a "true value" is problematic, they proposed it as a useful basis for the definition and then estimation of error. Their model is essentially a variance-covariance model and

permits considerable generalization (see for example Fellegi, 1964, 1974) and has been extremely influential among survey statisticians. In particular it allows the incorporation of the effects of interviewers and supervisors, and the possibility of correlated errors within households.

About the same time, Kish (1962) presented findings using an alternative technical approach using analysis of variance (ANOVA) models in dealing with interviewer error; this was the approach favoured by Yates, among others, and derived from the experimental design perspective of agricultural statisticians. Again the statistician simplifies reality so that it can be accommodated within the structure of his/her models; the effect of the interviewer is seen as an additive effect to the response of each respondent interviewed by that interviewer. The approach is easily generalizable to the effects of other agents in the survey (coders, supervisors, editors, etc.; see for instance Hartley and Rao, 1978); one drawback is that the ANOVA models do not easily lend themselves to the analysis of categorical data.

Figure 1.3 about here

These two approaches have in common the objective of estimating the variance of the mean of a single variable (or proportion). The focus of a survey is seen as the estimation of some important descriptive feature of the population. Thus as figure 1.3 illustrates, the total variance is seen as the sum of the various sources of error affecting the estimate. Starting with the variance of the sample mean of a variable measured without error and based on a simple random sample (SRS), a set of additional components may easily be added to the variance, each representing a separate source. Thus processing, nonresponse, noncoverage, and measurement errors can all be incorporated in the approach. For generality any biases - whatever their sources - may also be added in, giving the mean squared error as the total error of the estimate. This concentration on estimates of a single descriptive parameter arose partly from the government statistics orientation (which was frequently concerned with estimating a single important characteristic of the population, such as the unemployment rate) and partly from the general statistics tradition of interval estimation of the parameters of a distribution.

The survey statistics approach remained for two decades directed at such descriptive parameters. In the 1980s, however, the use of statistical models in data analysis and the controversy in survey sampling between the proponents of design-based and model-based inference led to the incorporation of response errors directly into statistical models in surveys. O'Muircheartaigh and Wiggins (1981) modelled interviewer effects explicitly in a loglinear analysis of the effects of aircraft noise on annoyance; Aitkin, Anderson and Hinde (1981) and Anderson and Aitkin (1985) used variance component models (also known as multi-level models) to investigate interviewer variability; Hox, de Leeuw, and Kreft (1991), Pannekoek (1991), and Wiggins, Longford and O'Muircheartaigh (1992) provide applications and extensions of these methods. Chapter 10 of this dissertation presents some recent work that extends the methodology.

These more recent developments mark a change in the statistical orientation from the original survey/statistical view of measurement error as a component of the "accuracy" of an estimate to a broader concern with the way in which measurement error can have an effect on substantive analysis of survey data. This latter view was of course always present in the case of the other survey traditions.

1.4 The Task

The task represents the core of the inquiry and the circumstances under which it is conducted. At its centre is the *questionnaire* or *interview schedule*. The exact nature and function of the questionnaire was not by any means universally agreed upon in the early years of survey research (nor indeed is it now).

Among the early exponents, Galton - not best known for his views on social surveys - set out in *Inquiries into the Human Faculty* (1883) his four requirements for questionnaires (see Ruckmick, 1930); they should"... *(a) be quickly and easily understood; (b) admit of easy reply; (c) cover the ground of enquiry; (d) tempt the co-respondents to write freely in fuller explanation of their replies and cognate topics as well ... . These separate letters have proved*

more instructing and interesting by far than the replies to the set questions". Booth (1889), a social reformer, took an instrumental view of the problem of data collection. *"The root idea ... every fact I needed was known to someone, and ... the information had simply to be collected and put together".*

In due course, however, social researchers began to question the reliability and validity of their data, and in some cases began to carry out experiments or other investigations to test and evaluate their methods. Psychologists had of course been aware of the possible effects of changing question wording; Muscio (1917) who describes experiments on question form refers to research by Lipmann published in 1907. One of the earliest directly related to the social survey was reported by Hobson in the Journal of the American Statistical Association in 1916 and contrasted different questionnaire types and wording in mail surveys. The bulk of research on this area did not emerge until the 1930s, with Link and Gallup to the fore, and the 1940s, which saw a dramatic increase in published work on methodology.

In this area of endeavour also there was evidence of the different perspectives of practitioners. From the psychologists came scaling methods, in particular the development of formal attitude scales (Thurstone, Likert, Guttman) spanning the period from 1920 to 1950. They brought with them their own terminology, using the terms reliability and validity to describe the characteristics of their measures. These terms have reassuring connotations of science about them, and they also emphasise the positive rather than the negative aspects of the measures. They may be contrasted with the statistician's *variance* and *bias* - terms with broadly similar meanings but emphasising the imperfections rather than the strengths of estimators. The psychological tradition also stressed the possibility of experimentation and experimental comparisons, often - as in the construction of attitude scales - using internal consistency, split-half experiments, and test-retest comparisons to examine their instruments. The Psychological Corporation was founded by Cattell to put applied psychology on a business footing; by 1923 nearly half the members of the American Psychological Association were stockholders (see Converse, 1986, p 107). One of the foremost practitioners in methodological (and applied) research was Henry Link, who joined the Psychological Corporation in 1930, and produced a succession of insightful papers on survey

and scaling methodology over the next twenty years.

On the academic front, the Office of Public Opinion Research at Princeton University was established in 1940 for the purpose of "*(1) studying techniques of public opinion research; (2) gaining some insight into the psychological problems of public opinion motivation; and (3) building up archives of public opinion data for the use of qualified students*" (Cantril, 1944, p. *x*). Among the task-related topics considered by Cantril and his colleagues were question meaning, question wording, the measurement of intensity (as distinct from direction) of opinion, and the use of batteries of questions.

On the whole, commercial market research practitioners were serious researchers into their own techniques. Cantril had forged a valuable alliance with Gallup - considered by many the father of the opinion poll - and his American and British Institutes of Public Opinion. During the period 1936 to 1949 Gallup conducted almost 400 split-ballot experiments. Many of these were designed by Cantril, to whom he gave access "*without restrictions or stipulations concerning the publication of results*". The major fruit of this effort was *Gauging Public Opinion* (1944) a compendium of studies of the methods of survey design, execution and analysis. It is a pity that there have not been more examples of such cross-sector collaboration in the fifty years since.

The other major work of that period arose from another exceptional (in this cased forced) cross-sector partnership. During the second world war a Research Branch had been set up in the Division of Morale, U.S. Army, directed by Stouffer (1941-45) (see Converse, 1986, p 165 *passim* for an excellent description). The war brought about a sustained programme of opinion research on matters of interest to the military establishment. The research was not conducted with any intention of furthering social science; the branch was set up to do a practical job by providing information to the authorities about the attitudes and views of military personnel. Staff and consultants of the Research Branch would produce a few years after the war the four volume work *Studies in Social Psychology in World War II* (1949-50) after a considerable amount of additional effort had been devoted to the material collected during the war. Volumes 1 and 2, together entitled *The American Soldier*, are a shining

example of how findings of methodological and theoretical interest can be found through appropriate analysis of routine applied work.

Throughout the fifties and sixties studies were published which illustrated various strengths and weaknesses of survey instruments (see for instance Parry and Crossley, 1950; Cannell and Fowler, 1963 as examples of validation studies). Sudman and Bradburn (1974) provided an overview of research to date and formulated an explicit model of the survey response process, the diagrammatic form of which is given in figure 1.1. Schuman and Presser (1981) published a valuable book describing their own and others' research on question wording.

Studies bearing on the survey task can be classified according to whether they used *validation information* or *internal consistency analysis*. In the former case - much the less frequent - information was available external to the survey that permitted checking the validity of the individual (or group) responses to the survey questions. Such information is hardly ever available in a substantive survey, and is in any case restricted to behavioral information. Methodological validation studies may themselves be considered as falling in two categories - *identification* studies where the objective is to identify and quantify errors of particular kinds, and *diagnostic* studies where the objective is to discover the factors generating a particular error and to devise an improved procedure to eliminate or reduce it (see O'Muircheartaigh, 1977).

The principal characteristics of the task that have over the years been identified as having a potential effect on the magnitude of the measurement errors include: the location of the interview and method of administration; the designated respondent; the length of the questionnaire; the position and structure of the questions; the question wording; the length and difficulty of the questions; the saliency of the questions; and the level of threat and possibility of socially desirable response. Other factors examined or postulated to have an effect were respondent burden, memory effects, the classic open vs closed question debate, the mode of data collection, explicit *don't know* category, number of points on a scale, and many more.

The difficulty with the literature on task variables is that no clear pattern was found to explain the many and various effects that were demonstrated to exist. The lack of a clear theoretical framework made it difficult to classify effects in any parsimonious way and hindered attempts to formulate a general theory of survey methods. While statisticians were content simply to quantify errors (and add their effects to the total variance or mean squared error), no satisfactory psychological, sociological, or practical principles were found that underpinned the miscellany of effects observed. Recognising the extent to which responses, and hence response effects, could be context dependent, it was suggested (Cantril, 1944, p. 49) that any result should be replicated across a variety of contexts: "*Since any single opinion datum is meaningful only in so far as it is related to a larger personal and social context, it is essential that responses to many single questions asked by the polls be compared with responses to other questions which place the same issue in different contingencies.*" This recognition of the context-dependent nature of survey responses may be seen as a precursor of later attempts to systematize our understanding of question-related effects.

1.5 The Interviewer and the Respondent

It is impossible to separate entirely the function and behaviour of the interviewer from the function and behaviour of the respondent in social surveys. This section describes how their roles developed and changed as the social survey and, more particularly, the way we think about the social survey changed and changed again.

The nature of data collection in surveys was by no means standardized in the early days of social investigations. In the case of many of the classical poverty studies there was no direct communication with the individuals about whom data were being collected. For Booth, for instance, "*the facts existed*" and all that remained was finding an efficient method of collecting them. He consequently used "expert" informants, such as School Attendance Officers, to provide detailed knowledge about children, and their parents, and their living conditions; Beatrice Webb later termed this procedure "*the method of wholesale interviewing*"; in general there was no concern about the performance of the interviewers.

The practice of having the interviewer record information without reference to the respondent continued, and in a limited context still continues for some kinds of information; Yates (1949 and subsequent editions) states: "*Thus it is better to inspect a house to see if it shows signs of damp than to ask the occupant if it is damp.*" DuBois, in Philadelphia in 1896, used a variety of methods, including a house to house canvass, to collect data on the black population of the city (DuBois, 1899). Rowntree (1902), a decade later than Booth, obtained his information directly from families by using interviewers.

There was also considerable variation in *how* interviewers were used. There were two poles to the interviewing role. One pole was represented by the expert interviewer who obtained information by having a "conversation" with the respondent, usually without taking notes at the time. The other was the "questionnaire" interviewer, who had a blank form and a prepared set of questions. Among the social policy reformers there tended to be some formal instruments in all cases, either a schedule of questions (leading to the term *interview schedule* for the questionnaire and accompanying instructions used by an interviewer) or a scorecard or tabulation card (subsequently becoming the *questionnaire*). Among market and opinion researchers there was a similar dichotomy: expert interviewers tended to be used when dealing with business executives or merchants, whereas "questionnaire" interviewers dealt with the general public. The questionnaire interviewers were given a short set of preprinted questions and wrote down the responses - in the presence of the respondent - on blanks provided for this purpose.

Partly because of the vagueness of the definition of a social survey and the absence of a generally recognised set of standards, the social survey was by no means held in universally high regard in its early days. Thomas (1912) in the American Journal of Sociology, opined that "... *interviews in the main may be treated as a body of error to be used for purposes of comparison in future observations.*" Gillin (1915) in the Journal of the American Statistical Association, decried the tendency towards lack of quality control: "... *the survey is in danger of becoming a by-word and degenerating into a pleasant pastime for otherwise unoccupied people.*" The interviewer and his/her behaviour was always seen as central to the quality of the survey. As early as the 1920s investigations were carried out into possible contaminating

influences of the interviewers (see, for example, Rice, 1929). By the 1930s momentum was gathering behind attempts to standardize the interviewer's behaviour drawing in many cases on criticisms of interviewing in contexts such a job interviews (see Hovland and Wonderlic, 1939; Bingham and Moore, 1934). In the 1940s a more systematic examination of the effect of interviewers was undertaken. Katz (1942) compared white-collar and working class interviewers and found a conservative tendency in results from white-collar interviewers; Cantril and his colleagues considered issues of interviewer bias, reliability of interviewers' ratings, interviewers' training, and rapport (Cantril, 1944); Wechsler (1940) drew attention to the manner of delivery of interviewers: "*An interviewer may ask a question in belligerent, positive tones with the obvious inference that he won't take no for an answer. He (or she) may use cadences so gentle that the person being interviewed will hesitate to voice what may appear to be a dissenting voice ... . Individually such cases may seem trifling. Add them up.*"

Gradually considerable pressure grew to standardize interviewers' behaviour and the leading survey organizations responded. In Public Opinion Quarterly (POQ) in 1942, Williams described the *Basic Instructions for Interviewers* for the National Opinion Research Center (NORC). These instructions are a model of clarity in their intent and describe a very carefully standardized practice. The Central City Conference in 1946 included a discussion of interviewing in which Moloney described the ideal interviewer ("*a married woman, 37 years old, neither adverse to nor steamed up about politics, and able to understand and follow instructions*"); this stereotype has persisted a long time. Tamulonis in the same session described experiments she had conducted that showed that middle of the road interviewers were less likely to bias their respondents than interviewers holding an extreme view on a question. Sheatsley (1947-48) adds advice on the use of interviewer report forms which presages the method of interaction coding proposed by Cannell in the 1980s.

The 1950s saw the publication of two major books on interviewing, by Hyman (1954) and Kahn and Cannell (1957). By this time the fundamentals of survey interviewing had been consolidated; interviewers were expected to behave as neutral instruments - professionals who read their questions as written without implying any favouritism towards any particular answer or expressing surprise or pleasure at a particular response. The ideal interviewer

would have a highly structured role, reinforced by rigorous training, and with no extra-role characteristics that might be salient for the questions being asked.

The role of the research subject (later the respondent) during this time had ranged from unimportant to insignificant. There was always concern about how information could be obtained. Booth, with his method of "wholesale interviewing" did not consult the individuals about whom he desired information. Though Rowntree and DuBois did ask the questions directly, there remained a general feeling that the respondent was something of an obstacle to effective research. There were seen to be problems with collecting data, but these problems were not *centred* on the respondent. Occasionally there would be concern that a task was simply too long and tedious. White, in 1936, wrote of the Federal Survey of Consumer Purchases that "*the respondent is supposed to know how many quarts of oil which he has bought during the year and have an account of the money he has spent for bridge and ferry tolls*"; White felt that anyone who actually kept such records "*would be so abnormal that their reports might not be representative of the public in general.*" On the whole, however, neither the respondent nor respondent fatigue was seen as a problem; the proceedings of the Central City Conference in 1946 sums up the lack of concern about respondents (and possibly sexism) in the observation "*They'd rather give their opinions than do their washing*"!

Even among thoughtful methodologists, the general view of the respondent was of a relatively passive actor in the research process; see for instance Sudman and Bradburn (1974): "*The primary demand of the respondent's role is that he answer the interviewer's questions.*" The main concern was with motivating the respondent to do so. Gates and Rissland (1923) were among the first to formalize this issue. Cannell (1975) reinforces the importance of motivation.

The influences that dictated the view of the interviewer and respondent roles described above were primarily the production requirements of large scale surveys. Survey methodologists had however been concerned about the possible shortcomings of the method. Cannell and Kahn (1968), in a review article, point out that the *interaction* between the respondent and

interviewer is central to the interview. Since then a series of studies has revealed an increasingly complex view of the interaction that takes place in an interview.

The studies by Cannell and his colleagues (summarized in Cannell et al., 1981) led to an explicit involvement of the respondent in the interview (through, e.g., commitment procedures) and the devising of feedback from interviewers that reinforced appropriate role behaviour by respondents. The group also developed behaviour coding schemes for interactions in the interview, leading eventually to a sophisticated instrument for pretesting survey questions and questionnaires (Cannell et al., 1989).

Rapport had always been seen as a key feature of a successful interview; the term encompassed a variety of qualities that implied success on the part of the interviewer in generating satisfactory motivation for the respondent. In the recent literature the concept of rapport has been replaced with "interviewing style", with the distinction made between formal or professional style on the one hand, and informal, interpersonal, socio-emotional, or personal styles on the other (see Schaeffer, 1991). Dijkstra and van der Zouwen (1982) argue that "*researchers need a theory of the processes which affect the response behaviour*"; the detailed study of the interactions that take place during the interview is one way of developing such a theory.

There had, of course, always been critics of the standardized interview; the criticisms were similar to those initially directed at the self-completion questionnaire. Even as, or perhaps because, surveys became increasingly used in social research, these criticisms surfaced again. Cicourel (1964) argued that there was no necessary relation between the researcher's agenda and the lives of the respondents. This point of view emerges again in the literature that considers the survey interview from the respondent's perspective.

The more qualitative approaches to the analysis of the survey interview tend to concentrate on the role of the respondent, or at least give equal status to the respondent in considering the issue. Suchman and Jordan (1990), in an analysis of a small number of interviews, describe the various and very different images that respondents may have of the survey interview,

ranging from an interrogation or test through to conversation and even therapy. The contribution of cognitive psychology can be seen in Tourangeau and Rasinski (1988) who provide a model of the response process which develops Cannell's (1975) description. This model of the process makes explicit some of the ways in which the active participation of the respondent is necessary if there is to be any chance of obtaining reliable and valid information. Even the rather esoteric field of "conversational analysis" has shown potential to contribute to our understanding of the survey response process (see Schaeffer, 1991).

There is a common thread that connects the contribution of all these sources. Overall the emphasis has shifted from the interviewer as the sole focus of attention, acting as a neutral agent, following instructions uniformly regardless of the particular respondent, to a joint consideration of the interviewer and the respondent taking part in a "speech event", a communication, where each has expectations as well as responsibilities. One pole of current thought is that of empowering the respondent (at its most extreme allowing the respondent to set the research agenda); the other pole is the ever more detailed specification of permissible interactions, so that while acknowledging the importance of respondent behaviour, the intention is to control it rather than to liberate it. These two poles are strangely similar to the original situation one hundred years ago, when the expert interviewer and the questionnaire interviewer represented the two extremes of data collection.

1.6 Recent Developments

A relatively theoretical paper in survey sampling in 1955 was instrumental (at least in part) in setting in train a profound change in the survey statistician's view of his/her role and the most appropriate ways to accommodate error in the statistical analysis of survey data. Godambe (1955) demonstrated the fact (neither surprising nor remarkable in itself) that there is no uniformly best estimator (even for the population mean) for *all* finite populations. This led, starting in the mid-1960s and lasting some twenty years, to a burst of interest in what became known as model-based estimation, i.e., analysis based not on the sample design (the joint probabilities of selection of the elements) but on some formal model that related the different variables in the analysis to each other, regardless of the configuration of the sample.

Smith's (1976) paper marks an important stage in that development.

This tendency was strengthened by the substantive demands of economics, psychology, and sociology, which lead to relatively complex analyses. The statistical methods used for these analyses are usually (or at least frequently) based on a *generalized linear model*; such models almost always assume *i.i.d.* (independent identically distributed) observations and often assume normality for the distribution of errors. The emphasis is on model specification and testing.

In contrast, the survey statistician's approach has traditionally been design-based. As section 1.3 illustrates, the primary concern is with inference to the particular finite population from which the sample was selected; response errors are conceptualized as another distortion of complete enumeration. In consequence weights are used to adjust the estimates in the light of the sample design and nonresponse, and standard errors are computed taking into account the effect of clustering of the sample and the correlated effect of interviewers and other agents. These adjusted standard errors are then used mainly to present confidence intervals for descriptive statistics or to test hypotheses about such statistics.

This difference of approaches illustrates the different emphasis placed on randomization and control by survey statisticians. For the survey statistician the overriding criterion has been randomization for *representation* through the sample design; for experimental statisticians and substantive analysts the emphasis has been on randomization and modelling for *control* of extraneous variables.

Recent developments in statistical methodology, allied to the dramatic increase in the power and accessibility of computers, have provided hope that these conflicting approaches may be reconciled. *Multilevel modelling* (a development of analysis of variance models; Bryk and Raudenbush, 1992; Goldstein, 1995; Longford, 1993), were originally designed to analyze hierarchical data structures such as education systems - schools, teachers, classes, pupils. This framework has been adapted to deal with the clustering imposed on survey observations by, inter alia, the sample design, response error, and repeated observations on the same

individuals. The analysis incorporates these elements explicitly in the analysis rather than ignoring them - as would be the case for standard statistical analysis - or adding them on as adjustments - which would be the effect of calculating design effects or interviewer effects for the standard errors. *Latent variable modelling* (Joreskog, 1973; Joreskog and Sorbom, 1989; Bartholomew, 1987) has its origins in the scaling literature in psychometrics. In such analyses the variance-covariance structure of the data and the errors is modelled explicitly and simultaneously. There are nontrivial challenges in formulating appropriate assumptions for each context, but there is no doubt that the approach has considerable potential.

1.7 Plan of the Dissertation

The dissertation considers response variance in sample surveys in the broader context of survey quality and survey error. Chapter 2 gives a brief overview of earlier research in the area. The principal content of the dissertation draws on investigations carried out by the author over the last thirty years. Three separate strands are identified: (a) simple response variance as the core component of response error; (b) interviewer variance as a component of total variance; and (c) the incorporation of errors directly into the analysis of the data.

Paradoxically the presentation of these strands does not follow what would appear to be the logical path from (a) through (b) to (c). The dominant approach, based as it is on the methods of the analysis of variance and ideas of interpenetration, is that represented by (b); the dissertation chooses to treat this first, in chapters 3, 4, and 5. The challenges offered by practical constraints for some major data sets give rise to a more detailed investigation of the core simple response variance ((a)) and this is dealt with in chapters 6 and 8; in an aside chapter 7 deals with interviewer assessment of responses and the variance of the variance estimates. Modelling (an element of (c)) arises both from (b) and (a), and this is presented in chapters 9 and 10. The final chapter provides an overview of the results and presents some arguments for re-conceptualizing error in survey data.

Chapter 2: THE TRADITIONAL SURVEY SAMPLING APPROACH

The dominant approach to accuracy and precision in the survey literature is that based on the mathematical analysis of the data in order to identify and estimate bias and variance components that contribute to the mean square error of survey estimates; this is the *official statistics* tradition described in section 1.3 of chapter 1. In this chapter the framework for that approach is set out and some early studies in this tradition are described, and the standard mathematical model for response variance is presented.

2.1 Introduction

Even in the case of a complete enumeration of the population the data, and the conclusions we reach, may be subject to serious errors due to faults in the method of measurement or observation. These *response errors* may arise from the questionnaire, from the execution of the fieldwork or from the nature of the data collection process. The form, extent, sources and effects of these errors are the concern not only of survey design but also of survey analysis. It may not be possible to eliminate such errors but it is possible to reduce their impact, estimate their effects and, in some cases, make use of them in the analysis.

The traditional survey approach bases its definitions of error on the concept of a 'true value' for each individual in the population. This true value must be independent of the conditions under which the survey takes place, which can affect the individual's response. The concept of *individual true value* of a variable for a population element was developed by Hansen, Hurwitz and Madow (1953) as follows:

(1) The true value must be *uniquely* defined.

(2) The true value must be defined in such a manner that the purposes of the survey are met.

(3) Where it is possible to do so consistently with the first two criteria, the true value should be defined in terms of operations which can actually be carried through (even though it might be difficult or expensive to perform the operations).

It is possible to define the true value in such a way that there are no response errors. A respondent's age could be defined as the answer given to the question 'What is your age?'. Similarly a respondent's attitude to the reintroduction of the death penalty could be defined as the answer s/he gives to the question 'Do you think the death penalty should be reintroduced?' Both definitions satisfy the first and third criteria: the response is unique and it is defined in terms of operations which can be carried through. However, it is probable that they do not satisfy the survey objectives in such a way that such 'true values', dependent as they are on the specific conditions obtaining at the interview, would be acceptable as an ideal, although they might be acceptable as approximations to the true value. The individual true value should be seen as a characteristic which is independent of the survey conditions which affect the individual response. Age, for example, is defined as a time interval between two events, and this definition is independent of the method by which, and the conditions under which, we determine or observe the individual's age. For some other variables, such as income, the true value may be easy to define but difficult to obtain. For attitudinal items even the definition of the true value may be obscure. In all cases however the individual true value is a useful ideal at which to aim and the consideration of departures from this value is helpful in assessing the methods by which we obtain information.

The term *individual response error* is used to denote the difference between the individual true value and the observation recorded for the individual. For example, if for a respondent born on January 16th 1946 age were recorded on January 16th 2000 as 51 years, the individual response error would be 3 years. The individual response is defined as the value obtained for a particular observation. Under different conditions (with a different interviewer or with a different form of question, for instance) a different observation might be obtained.

It is useful to distinguish between two components of the response error. The distinction is based on the definition of some of the characteristics of a survey as the *essential survey conditions* (Hansen, Hurwitz and Bershad, 1961): for example, the subject matter, the data collection and recording methods, the timing and sponsorship, the type or class of interviewers and coders to be used in an interview survey, etc., can be considered as essential parts of the survey design. These characteristics may be seen as stable across a set of surveys

for a particualr survey organization, and as being difficult to change without a major change in the organization. The expected value under these conditions can be defined. The difference between this value and the true value is the *response bias*, either for the individual or for a group of individuals.

In addition to this there are 'random' fluctuations ("variable errors") about the expected value. The particular interviewers chosen from the designated class of interviewers, the particular coders, and transient characteristics of the observation situation are sources of such fluctuations. These variable errors also contribute to the response error, in the form of *response variance*. In order to appreciate the meaning of response variance it is necessary to postulate that a survey is conceptually repeatable under identical conditions, the essential survey conditions. A survey is then seen as a single observation from a set of possible observations. The response variance is a measure of the variability among these observations.

The response bias and response variance differ also in implications for the survey analyst. First, the bias term is a constant that cannot be measured from within the survey; it is necessary to have data from some other source in order to assess it. On the other hand, the different components of response variance could in principle be estimated from the survey observations themselves. Second, the effect of the response bias is fixed regardless of the number of observations taken. By definition even a complete enumeration would have the same response bias under the same essential survey conditions. The effect of response variance can, however, be changed by sampling a larger number of the units involved. By increasing the number of interviewers, for example, the interviewer variance can be reduced. Third, response bias is of particular concern in the estimation of means and totals for the whole sample. For comparisons between means of subclasses and in the calculation of measures of correlation and association the effect may be slight. Response variance will not affect the expected value of estimators of means and totals but will contribute to their imprecision. In addition, response variance will tend to attenuate measures of relationship between variables. This will apply even to the comparison of two subclass means if the classifying characteristic is subject to measurement error. These latter issues become even more important when we attempt to identify the impact on statistical analysis of response

variance in the data.

It is worth noting here, though I will not pursue the topic in this dissertation, that the definition of what constitutes the boundary between bias and variance is essentially determined by the level at which the definitions are conditioned. This is the role played by the essential survey conditions in the Hansen, Hurwitz and Bershad (US Census Bureau) framework. By changing the definition of the essential survey conditions we could transform a bias into a variance and vice versa; this would not of course mean that we could necessarily estimate these components of error.

2.2 The Identification and Diagnosis of Response Errors

2.2.1 The use of validation information

It is only from sources external to the survey that information that permits checking the validity of the individual results can be found. Such information is hardly ever available for data collected in a substantive survey. If the information does exist it can be used to identify the individual response errors for particular items. This approach is limited to behavioural information; individual true values cannot be obtained for attitudinal information. Predictive validity may be relevant for some attitudinal data, for instance the prediction of voting behaviour from stated voting intention. But failure of data to predict behaviour does not necessarily mean that the responses were not measuring the respondent's attitudes at the time at which they were recorded.

Methodological Studies

Methodological validation studies are here considered as falling into two groups: first, those that are concerned with estimating the magnitude of effects and identifying the areas (type of respondents or types of questions) where they occur; second, studies whose aim is the isolation of factors in the role performances of the interviewer and respondent for the purpose of improving fieldwork procedure and training in order to eliminate the errors. The first

group may be called *identification* studies, the second, *diagnostic.*

Identification Studies

Two early studies which demonstrated the existence of response errors for behavioural or factual data for which external sources of validation were available are described briefly here. The paper by Parry and Crossley (1950) describes a study which was designed to investigate the validity of responses to a set of questions which were thought to evoke varying amounts of prestige and varying degrees of potential distortion caused by social pressure, ease of verification, and memory factors. The eight items chosen were:

(1) Respondent's registration and voting in the six city-wide Denver elections from 1944 to 1948. The voting history given by the respondent was checked against official lists of voters.
(2) Personal contribution during Community Chest drive.
(3) Possession of valid Denver public library card in respondent's name.
(4) Possession of valid Colorado driver's licence.
(5) Ownership of a car by respondent or spouse and make and year of car.
(6) Respondent's age.
(7) Ownership or rental of respondent's place of residence.
(8) Telephone in respondent's home.

Items 2, 3, 4, 5, 7, and 8 were checked against various official records. Item 6, age, was checked against various lists. Since respondents could also report incorrectly to these lists this check was less satisfactory. The response rate was 68 per cent which was perhaps a little low, but does not affect the precision of the errors obtained for those who responded. The checking was a name-by-name manual operation.

More respondents over-reported participation in elections than under-reported. One third of all respondents reported contributing to the Community Chest where records indicated no contribution. The number of library cards and driver's licences were also exaggerated. These

results support the idea that the pressure on the respondents is towards claiming socially desirable attributes.

A similar study is reported by Weiss (1968). The purpose was to investigate whether respondents in lower socioeconomic groups deliberately misreported attitudes or behaviour. The procedure was to select a group of African-American 'welfare mothers', ask them a number of questions and then compare the answers with available official records. The questions dealt with (i) registration status and voting, (ii) receipt of money from welfare, (iii) children's educational performance. In addition to behaviour questions, attitudinal questions (which could not be verified from records) were also asked.

The questions on welfare were answered extremely accurately (> 98 per cent) perhaps because these questions followed a series of questions that assumed participation in the welfare system. For the questions dealing with children's educational performance there may have been considerable genuine lack of knowledge on the part of the mothers which could not be distinguished from 'bias' - here taken to mean deliberately reporting more socially desirable behaviour than the facts warranted. On the voting and registration questions, the amount and direction of response errors were similar to those of the largely middle-class population whose voting self-reports have been investigated in previous studies. As far as this bias is concerned two results emerged: (i) the greater the status similarity between the interviewer and the respondent the greater the response bias is likely to be; (ii) the more friendly and personal the interview, the greater the bias.

Diagnostic Studies

In diagnostic studies the emphasis is on identifying particular aspects of the interview situation in terms of the contribution they make to response errors. The effectiveness of different procedures can then be evaluated through the use of experimental studies utilizing the available validation information. The work can concentrate either on the task variables or on the role performances of the interviewer or respondent. Cannell and others at the Institute for Social Research of the University of Michigan (ISR) carried out a series of such studies

which are described briefly below. This was, and is, one of the most promising approaches to the investigation of response errors, concentrating as it does on the basic elements of the data collection process and having as its aim the reduction (and possible elimination) of response errors by means of improved role performance by the interviewer and respondent.

Method of obtaining information. Cannell and Fowler (1963) report a study dealing with hospitalization in which both a self-enumeration procedure and personal interviews can be compared with external validating data. From a probability sample of general hospitals, a probability sample of individuals was selected from discharge records. The names and addresses thus selected were assigned randomly to the two procedures. The interviewers were not aware that the study involved a record check. The following conclusions emerged:

(i) For questions where the information is most likely to be improved by checking records and consulting with other people, the responses are more accurate for the self-enumeration procedure. These are, for example, questions dealing with length of stay in hospital or date of discharge. For questions for which no records are available to the respondent, the answers to the personal interview are more accurate. This may be because the interviewer can better motivate the respondent to obtain more complete answers. Questions on diagnosis or form of treatment administered fall into this category.

(ii) When a three-point rating scale based on degree of social 'threat' or embarrassment thought to be connected with reporting a particular diagnosis was used to classify questions, no evidence was found to differentiate between the quality of responses for the two procedures. However, the respondent was not anonymous in either case. It is possible that anonymity, and not the presence or absence of an interviewer, is the important factor in such cases.

(iii) The results were compatible with the hypothesis that the information about persons who have proxy-respondents is reported more accurately in the self-enumerative procedure because of the opportunity to consult freely with other members of the family or even with the member about whom the information is being sought.

(iv) The results were compatible with the hypothesis that the motivational level of the

respondent has more effect on response error in the self-enumerative than in the interview procedure. 'Motivation' was measured by the extent to which the respondent volunteered information without being asked for it specifically. For the self-enumeration, it is indicated by the speed with which the respondent replies and whether reminders are needed.

(v) There was no evidence that the personal interview was superior for respondents with low level of education.

Interviewer performance over time. It is generally believed that experience leads to better performance by the interviewer. In fact many studies (e.g. Durbin and Stuart, 1951; Booker and David, 1952) have shown substantial differences between results for experienced and inexperienced interviewers. However, the performance of an interviewer during the course of a single study has been shown by Cannell *et al.* (1970) to deteriorate over time, both for experienced and inexperienced interviewers. Performance began to deteriorate immediately after training and in some cases dropped significantly in a few weeks. The responses were validated against official records and the deterioration was in the form of more severe under-reporting of hospitalization.

The hypothesis put forward to explain this deterioration is based on the *motivation* of the interviewer. The interviewer's *ability* to fulfil the role is not likely to change drastically in a short time, but motivation may decrease due to the lack of reinforcement of good performance during the course of the study. Performance, in the sense of eliciting complete reporting, cannot be judged on the basis of the completed questionnaire alone. In any case the results suggest that the interviewer's behaviour has an effect on the respondent's behaviour not only during the interview but also when the respondent is asked to fill out and mail a self-administered form.

Cannell suggests that there are two main implications of these findings. First, that there is a need to identify the elements of interviewer behaviour which are related to adequate reporting. Second, that supervisory and training procedures are needed to stimulate or reinforce adequate interviewer role performance during the course of the fieldwork itself.

Respondent Commitment. Cannell and his colleagues developed a commitment procedure for use in a health interview. The principles were based on the work in social psychology where Lewin, for example, demonstrated that people who made a public decision to do something were more likely to carry out their own decision than those who had not made a public commitment. The objective was to improve respondent reporting by getting a commitment to adequate role performance. Once a few questions had been asked, a statement was read to the respondent about the importance to the research of getting complete and accurate information which would require some diligence, and that it was important to report accurate information even though it might be embarrassing. The respondent was then given the choice of stopping the interview or of signing a statement of commitment to give accurate and complete answers. A statement guaranteeing anonymity was signed by the interviewer. Only 8 of the 192 respondents who were presented with the statement refused to sign. A control group, for which the questionnaire, interviewing techniques, and interviewers were the same, but with no commitment procedure, was interviewed in the same study.

The findings indicated that the commitment procedure produces more precise and more complete reporting of a variety of kinds of information. It appeared that commitment also encourages the reporting of potentially embarrassing information. Overall the procedure had the desirable effect of causing respondents to perform their tasks with more diligence than they otherwise would.

Feedback. The quantity and type of feedback given to the respondent has been shown to affect the respondent's reporting behaviour. The feedback can have two uses. First, to inform the respondent of what is expected of him (how to answer a question, what constitutes a satisfactory answer, etc.). Second, to motivate him to further effort. In general in survey interviews very little, if any, guidance is given to the interviewer on how to react to responses. In fact, some of the research at the Institute for Social Research (ISR) shows that 40 per cent of the verbal interchange in an interview includes interviewer and respondent behaviour over and above that of asking or answering the interview questions. In order to test the effectiveness of various feedback reactions, an experimental questionnaire was designed by Cannell and his colleagues which included explicit positive and negative reinforcement

statements. The purpose was also to test the feasibility of using such statements in training and structuring interviewer behaviour.

The final results of the analysis indicate that the reinforcement procedure tended to reduce recall errors but was not effective in improving responses on threatening or socially embarrassing questions.

Substantive Studies

Typically for substantive studies the opportunities for validating the responses through the use of external sources are very limited. It is usually not possible to obtain external information on all the elements in the sample. [Indeed, if it were, the sample survey would probably not have been necessary to obtain this information.] In some cases, cost is the principal consideration in carrying out a sample survey when accurate information is available elsewhere. In these cases, validating a sample of the responses may provide a guide to the magnitude of the errors involved. For example, in relation to the U.S. Population Census of 1950, Eckler and Pritzger (1951) reported the use of six record checks on a sample of the Census responses.

Checks of this kind have serious problems. First, it is normally not possible to locate the information even for all of the check samples. Second, the records used in the checking procedure may themselves be inaccurate. Third, when discrepancies occur between the records and the responses, without further fieldwork it may be impossible to reconcile the differences.

2.2.2 Internal consistency analysis

In most situations it is not possible to obtain information from external sources to enable direct validity checks to be carried out. However, much can be done by appropriate design of the survey and analysis of the responses themselves. Three such diagnostic studies are described below.

Method of collecting information. Hochstim (1967) describes a study designed to investigate the relative merits of mail questionnaires, telephone interviews and face-to-face interviews for two types of data.

The Human Population Laboratory of the California State Department of Public Health undertook a study based on three 'strategies' of data collection, each starting with one of the basic methods (mail, telephone, personal interview) and supplemented with other methods as needed (e.g. *mail strategy*; questionnaires sent, then first reminder to those who did not return the first questionnaire; if necessary a second reminder; those still not responding were called upon, by telephone, or in person if necessary). All other aspects, including the questionnaires, were to be held constant so that the means of collecting information could be evaluated by themselves.

Two studies were carried out in which the three strategies were tested:

(a) the Sampling Frame Study: questions dealing with medical, familial, behavioural, demographic characteristics (all occupiers of household aged 17 and over); and

(b) the Cervical Cytology Study (Papanicolaou cancer detection test) provided an opportunity to increase the influence of topic sensitivity on return rates (only women, 20 years and over). The data gathering techniques for the two studies were identical and everything possible was done to achieve a high rate of returns for all three strategies. All three strategies were based on area probability samples.

The responses from the three strategies were found to be highly comparable. The responses for each strategy were compared (a) with available census data; (b) in terms of number of statistically significant differences; (c) in terms of magnitude of actual percentage differences; and (d) in terms of content of questions on which major interstrategy differences occurred. On (a), (b) and (c) the strategies were practically interchangeable. Some sensitivity to particular topics was found in the differences between the strategies which was in line with other research in this area: the proportion of women saying they *never* drink, for example, was substantially higher when the respondent faced an interviewer than when she returned a

They used two kinds of recall procedures: (i) *unbounded recall*: respondents were asked to report expenditures since a given date and no control was exercised over the possibility that respondents might shift some expenditures into or out of the recall period; (ii) *bounded recall*: at the beginning of the interview, which must be the second or later interview within a household, the interviewer informed the respondent of the expenditures which had been reported during the previous interview and asked for additional expenditure since then. In order to study the effect of the length of the recall period both bounded and unbounded recall interviews were conducted with varying lengths for the reference period. By using an ingenious panel design involving fifteen pairs of samples, data for each month of the study were available for a variety of recall procedures, and comparison of estimates from different procedures for any given time period reflected the effect of the procedure only.

The unbounded recall of expenditure for the month preceding the interview was found to involve substantial net forward telescoping into the recall period. This effect was greater for larger jobs. For the three-month recall period, the largest part of the net *external telescoping* went into the earliest month of the recall period, i.e. into the month most distant from the time of interview. In this period there was substantial *internal* forward telescoping, i.e. there was a tendency to transfer expenditures to more recent points in the time period. It was found also that for the number of larger jobs, the reporting at the unbounded recall interviews was over 50 per cent higher than that at the bounded recall with a one-month recall period. With a three month unbounded recall, the rate was about 26 per cent higher than for bounded recall. Bounded recall procedures thus reduced the effect of forward telescoping and, when possible, reduced the error in the responses.

2.2.3 The factorial design

Durbin and Stuart (1951) used a factorial design to measure interviewer variability. The orientation of the work was towards comparing organizations (classes of interviewers) rather than individual interviewers: interviewers within organizations were considered to be homogeneous. The factors which were assessed in terms of their contribution to the total variance were organizations (α_i, 3 levels), questionnaires (β_j, 3 levels) districts (γ_k, 3 levels),

age of subject (δ_l, 4 levels), and sex of subject (η_m, 2 levels). These produced a 3×3×3×4×2 factorial design. There were seven replications for each factor combination (i.e. seven respondents in the sample for each of the 216 factor combinations). The model therefore was

$$y_{ijklmn} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \eta_m + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \\ \ldots + 3^{rd} \text{ and } 4^{th} \text{ order interactions} + (\alpha\beta\gamma\delta\eta)_{ijklm} \\ + \varepsilon_{ijklmn}$$

This model permits the estimation of very complex effects. For example it is possible to estimate the interaction $(\alpha\beta)_{ij}$ between class of interviewer (organization) and type of questionnaire.

Gales and Kendall (1957) used a similar design. There were four factors: organizations (6), questionnaires (2), briefings (2) and areas (4). A complete factorial would have involved 6×2×2×4 = 96 pairs of interviewers. Having only 24 pairs of interviewers available, some partial replication (and confounding) were necessary.

The main emphasis of the Durbin and Stuart paper was on the variation in success rates and non-contacts rates between the factor combinations. The analysis assumed that interviewers within organizations were homogeneous, after testing for homogeneity using a χ^2 test of homogeneity for a binomial series. They found highly significant differences between experienced and inexperienced interviewers for both successes and refusals. Significant differences were also found between questionnaire types and between districts (in the case of successes) and for age and sex of respondent (in the case of refusals). No significant effects for interactions involving interviewer organization were found. Their conclusion was that performance (success rate) may be regarded as the sum of independent effects, i.e. it may be explained by a simple additive model.

Gales and Kendall present results for differences between organizations, all of which had experienced interviewers. Significant differences were found on two questions, and a number of others, together with interaction terms (e.g. organization/questionnaire interaction), were

suggestive. These differences were interpreted as being due to interorganizational differences rather than differences between individual interviewers. The results provided support for the contention that interviewer effects are larger for attitudinal and ambiguous items than for clearly defined factual items. The study also indicated that type of briefing may contribute to interviewer effects, especially on questions which are not absolutely clear, and that type of question may also contribute, especially if asked in an open-ended form.

2.2.4 A simple comparative model

The Sudman and Bradburn (1974) model of the response process presented in chapter 1permits the comparison of response effects between studies and also between behavioural and attitudinal data. The *relative response effect* is defined as

$$RE = \frac{(\text{Actual response} - \text{Validating score})}{s}$$

where s is the standard deviation of the population, obtained from the validation information where possible. If no information is available on the size of s an estimate of RE can be made using RE = (Actual - Validating)/(Validating). For attitudinal information, the weighted mean of all observations was used for validation. This measure has a number of drawbacks, the most serious of which is that the size of the relative effect computed for a subgroup may depend on the size of the subgroup in the sample.

Using data from a large number of studies, the greater proportion of which were based on samples of American respondents, Sudman and Bradburn estimated response effects for a large number of factors and for interactions between factors. The principal results obtained were:

(i) For threatening questions, self-administered questionnaires performed better than face-to-face interviews. In addition, on questions involving social desirability, there was a greater tendency to conform in face-to-face interviews.

(ii) Large response effects were observed when college students, particularly males, were either respondents or interviewers.

(iii) The greater the saliency of the questions for the respondent, the lower the response effect.

(iv) For threatening questions, closed questions were worse than open-ended questions.

(v) Reports by others (proxy reports) were almost as accurate as self-reports; 'threat' had no influence on non-self reports. The issue of self vs proxy reports is discussed in detail in chapters 8 and 9.

(vi) Aided recall procedures (specific alternatives listed etc.) lead to an increase in response effects. The use of bounded recall procedures and the availability of records for checking both lead to a decrease in response effects.

(vii) Particular extra-role characteristics of the respondents and interviewers and the interaction between these and the interview conditions were very topic-specific and no general results emerged.

The approach adopted by Sudman and Bradburn is useful in that it provides some basis for the measurement and comparison of the magnitudes of response effects. But the model has many weaknesses and cannot control for many of the factors which may influence the direction and magnitude of response errors. One way to approach the problem is to identify and specify a functional form for the operation of response errors and to evaluate the parameters of the functional form. This is particularly important for attitudinal data where validation information will not exist. And it is also crucial if the effect of the errors on the estimation process is to be investigated.

2.3 The Survey Model

2.3.1 Introduction

It is important that the analysis of response errors should not lose relevance to real problems due to the sophistication of the mathematical approach. The basic objective of a survey is to provide data on the basis of which the survey variables can better be understood, described

or predicted. The aim in the analysis of response errors should therefore be to maximize the information which can be abstracted from the data. In the context of the most surveys, methodological experimentation is by and large precluded by the very nature of the operation. The primary objective is to obtain the best possible data from a single operation, which necessarily requires the choice of a study design considered a priori to be the most suitable. Thus it is not possible in general to compare different survey procedures in order to ascertain which is superior. Furthermore for almost all survey data, there is no source of external validation data available at the level of the individual respondent. Consequently the analysis of response errors must generally be based on an examination of the internal consistency of the data. The emphasis is on the *reliability* of the data, rather than on its validity; in other words, on response variance rather than on response bias.

A severe constraint is therefore imposed on any large scale investigation of response errors in surveys. It is not possible to interfere with the principles laid down for the conduct of the survey by introducing any new or experimental procedures which *might* reduce the quality of the data collected. Furthermore it is not possible, given the absence of an external source of validation data, to examine the absolute magnitudes of the individual response errors. There are however two possible approaches which can provide some information on the magnitude and impact of the errors; *re-enumeration* and *interpenetration.*

The first approach, re-enumeration, involves re-interviewing at least some of the respondents in the main survey. The re-interviews should be carried out soon after the main survey under the same (or similar) essential survey conditions. This would provide two separate observations on each of these respondents.

Certain characteristics of the survey would be constant for the two observations: the subject matter, the questions asked, the field force, the procedures for the supervision and control of the fieldwork, the coding and processing of the questionnaires. Thus the data could provide no information on the effects of these conditions on the survey results. In order to assess the systematic impact of any or all of these factors, either some source of information outside the survey procedure or an experimental design controlling these factors would be necessary.

Some factors would vary between the two surveys, however. The transient situational factors certainly vary, the two interviews being conducted on different occasions in every case. In addition, two different interviewers could be used for each individual and thus a part of the difference between the observations might be due to differences between the interviewers. The same would be true of the coding and processing, although the allocation of questionnaires to coders might not be conducted as rigorously as the allocation of respondents to interviewers.

In essence, therefore, such data would not, and could not, provide any information on response bias. Without external validation data, no assessment can be made of any *systematic* distortion of the observations produced by the conduct of the survey. What they would provide is an opportunity to examine the *reliability* of the measurements, the extent to which the application of the same essential survey conditions on two occasions would produce different results. Thus, they would afford an opportunity to partition the variability observed in the survey observations into two components, one due to the inherent variability in the variable being measured, the other introduced into the recorded responses by the observation process itself.

The second approach, that of interpenetration, involves a modification of the survey design. It has been established in other contexts that interviewers may influence in a systematic way the responses they obtain. When this is so, the estimates of variability obtained in the usual way for statistics calculated from the sample observations may seriously underestimate the true variance. This component of variance - the *correlated response variance* due to interviewers - will be present in any statistics calculated from the survey data, but the difficulty in practice is that there is usually no way of estimating it. The problem arises because respondents are usually allocated purposively (or haphazardly) to interviewers and any differences between the results obtained by different interviewers may be due to differences between the groups of individuals whom they interview rather than to differences in the influence of the interviewers themselves. It is possible, however, to modify the survey execution in such a way that this component of variance is estimable. The basic feature of the design is that (at least within certain defined limits) the respondents must be allocated

randomly to interviewers, so that no systematic difference between the workloads of the interviewers should contaminate the comparison of the results of the interviewers. There will of course be differences among the workloads, but as long as the allocation of respondents to interviewers is random, these differences can be taken into account in the analysis. This procedure of random allocation of workloads is called *interpenetration*, and was the cornerstone of Mahalanobis' work in India on response variance.

It is obviously impossible in practice to allocate a random subsample of a national sample to each interviewer. Not only would the cost of such an operation be enormous, but the disruption of the field execution of the survey would make it unacceptable in terms of the objectives. However, the field strategy of a survey may lend itself to a modification of the design which is equally satisfactory. In the World Fertility Survey (WFS), for example, interviewers worked in teams, a team usually consisting of four to six interviewers and two supervisors responsible for organizational supervision and timely scrutiny of interviewers' work. Each team worked and travelled as a unit. The allocation of work to the interviewers was normally the responsibility of the supervisors. The supervisors had, for each area, a list of the individuals (or in some cases, households) to be interviewed. It would obviously be a straightforward matter to determine the allocation of respondents to interviewers before the fieldwork in such a way that each interviewer is allocated, in effect, a random subsample of the work in that area. Similarly, in telephone surveys, workloads may be randomised within shifts in the centralised interviewing facility, and in localised surveys complete interpenetration may be used.

Thus, without any significant interference with the procedures of data collection, it is possible to modify the execution of the survey so that the contribution of the correlated response variance due to interviewers could be estimated and its impact on the survey results assessed.

The basic approach thus involves two elements:

1 *Re-enumeration* At least a subsample of the respondents in the main survey can be re-interviewed under the same (or similar) essential survey conditions. This will

permit the partitioning of the observed variability of the responses into two components: the sampling variance and the simple response variance. It will also make it possible to examine in detail the extent to which the same individuals (the respondents) give identical (or different) answers to the same questions on different occasions.

2 *Interpenetration* By allocating interviewers' workloads randomly (perhaps within teams), it will be possible to estimate the extent to which the usual estimates of variance underestimate the true variance, and thus to provide a more valid estimate of the total variance of the survey.

In some cases the particular design of a project may combine the two procedures of interpenetration and re-enumeration in a way that permits the estimation of some additional parameters of the response errors.

There are three conceptually distinct sources of variation in the results from a survey.

1 The variation among the true values for different individuals. These true values are the quantities of interest in the survey itself. The true value for each individual is fixed. The only variability to which the results would be subject if the true values were observed directly would arise from the fact that typically only a sample from the population is observed. The sampling variance of the estimator is the variance of the sampling distribution of the estimator and depends only on the sample design.

2 The value of an observation is determined not only by the true value for the individual but also by errors of measurement. The sources of these errors are many and their impact will vary considerably from one variable to another. We can specify a general form for the distribution of these errors even though we do not know all the particular influences which generate them. In order to specify such a distribution, we need to assume that a survey is conceptually repeatable. The distribution of the response errors is identifiable only from replications of the survey. If we can, however, identify

particular potential sources of response errors it is possible to obtain a measure of their impact by appropriate design of the survey.

3 The third potential source of variation in the observations arises from possible interaction between the observation process and the true values of the individuals in the sample. Since in the theoretical treatment we assume that we are dealing with a simple random sample (SRS) of individuals from the population, the only way in which this may come about is that the allocation of the sample elements among interviewers influences the response errors generated by the interviewer. Thus it is conceivable that the particular sub-sample allocated to one interviewer may affect the response errors within that interviewer's workload. The probability distribution of these components of error is therefore over different samples and different allocations of the sample. Where the sample design is not SRS there are other circumstances in which there may be a covariance between the sampling variance and the response variance. These are considered in chapters 3 and 11.

2.3.2 The mathematical model

The basic mathematical model for response errors is given below. The model is applicable to all variables which are measured on an interval scale, and also to binary variables. For simplicity the discussion is restricted to the estimation of the population mean. This is the *Census Bureau Model*, developed by Hansen, Hurwitz, and Bershad (1961), and extended by Fellegi (1964). Some minor modifications to the estimation procedure are presented in cases II and IV below.

A particular survey is regarded as a single trial, ie the survey is regarded as conceptually repeatable. An observation for the jth element in the survey for trial t is denoted by $y_{jt(s)}$ where j denotes the individual, t denotes the trial and s denotes the sample and its allocation.

The observation $y_{jt(s)}$ can be partitioned as follows:

$$\bar{y} = \sum_{j=1}^{N} y_j$$

In assessing the impact of response errors on the estimator, it would be possible in principle to study any sample design. The presentation below is confined to simple random sampling. The sample mean of the observations is

$$\bar{y}_{.t(s)} = \frac{1}{n}\sum_{j\in s} y_{jt(s)}$$

Define $\Delta_j = y_j - \bar{y}$. The observation $y_{jt(s)}$ may now be written as

$$y_{jt(s)} = \bar{y} + \bar{\Delta}_{.(s)} + \bar{\beta}_{.(s)} + \bar{\varepsilon}_{.t(s)} \qquad (2.3)$$

where

$$\bar{\Delta}_{.(s)} = \frac{1}{n}\sum_{j\in s} \Delta_j$$

$$\bar{\beta}_{.(s)} = \frac{1}{n}\sum_{j\in s} \beta_{j(s)}$$

$$\bar{\varepsilon}_{.t(s)} = \frac{1}{n}\sum_{j\in s} \varepsilon_{jt(s)}$$

The difference (net deviation) between $\bar{y}_{.t(s)}$ and $\bar{y}$ is made up of three components: the sampling deviation, the fixed response deviation and the average variable response deviation.

The overall bias of the strategy is given by

$$\begin{aligned} E(\bar{y}_{.t(s)} - \bar{y}) &= E_p E_\eta (\bar{y}_{.t(s)} - \bar{y}) \\ &= E_p(\bar{\Delta}_{.(s)} + \bar{\beta}_{.(s)}) \\ &= E_p (\bar{\beta}_{.(s)}) \end{aligned}$$

Thus the overall bias is due only to the fixed response effects. This *response bias* is not estimable from the survey data themselves.

The mean squared error of $\bar{y}_{.\pi(s)}$ can be obtained from (2.3) and is

$$E_p E_\eta (\bar{y}_{.\pi(s)} - \bar{y})^2 = V_p (\bar{\mathcal{A}}_{.(s)}) + V_p (\bar{\beta}_{.(s)}) + E_p E_\eta (\bar{\varepsilon}^2_{.\pi(s)}) + 2\mathrm{cov}_{p\eta} (\bar{\mathcal{A}}_{.(s)}, \bar{\varepsilon}_{.\pi(s)}) + 2\mathrm{cov}_p (\bar{\mathcal{A}}_{.(s)}, \bar{\beta}_{.(s)}) + [E_p (\bar{\beta}_{.(s)})]^2$$

The total variance of $\bar{y}_{.\pi(s)}$ is given by

$$E_p E_\eta [\bar{y}_{.\pi(s)} - E_p E_\eta (\bar{y}_{.\pi(s)})]^2 = V_p (\bar{\mathcal{A}}_{.(s)}) + V_p (\bar{\beta}_{.(s)}) + E_p E_\eta (\bar{\varepsilon}^2_{.\pi(s)}) + 2\mathrm{cov}_{p\eta} (\bar{\mathcal{A}}_{.(s)}, \bar{\varepsilon}_{.\pi(s)}) + 2\mathrm{cov}_p (\bar{\mathcal{A}}_{.(s)}, \bar{\beta}_{.(s)}) \quad (2.4)$$

The overall variance of the estimator can therefore be partitioned into four components, ie

overall variance =

sampling variance (of true values)
+ variability due to fixed response errors
+ variability due to variable response errors
+ covariance between sampling deviations and response effects.

The model whose total variance is given by (2.4) is the general model for the system. In the subsections below some special cases are considered.

Case I Simple Response Variance

The simplest situation is that in which the only distortion of the true values is a random disturbance term which is uncorrelated with the true values. This can be expressed by modifying the model (2.1) and (2.2), using the following simplifying assumptions:

$$\beta_{j(s)} = \beta, \text{ all } j$$

$$V_\eta(\varepsilon_{jt(s)}) = \sigma_j^2 = \sigma_\varepsilon^2 \text{ , all } j \tag{2.5}$$

$$\text{Cov}_\eta(\varepsilon_{jt(s)}\varepsilon_{j't(s)}) = \rho_{jj'}\sigma_{j(s)}\sigma_{j'(s)} = 0 \text{ all } j,j'$$

and the total variance (2.4) becomes

$$E_p\, E_\eta \left[\bar{y}_{.t} - E_p\, E_\eta(\bar{y}_{.t})\right]^2 = V_p\left(\bar{\Delta}_.\right) + E_p\, E_\eta\left(\bar{\varepsilon}_{.t}^2\right)$$

$$= \left(1 - \frac{n-1}{N-1}\right)\frac{\sigma_y^2}{n} + \frac{\sigma_\varepsilon^2}{n}$$

If the finite population correction is ignored, this gives

$$V_{p\eta}\left(\bar{y}_{.t}\right) = \frac{\sigma_y^2}{n} + \frac{\sigma_\varepsilon^2}{n} \tag{2.7}$$

The first term in (2.7) is the *sampling variance* of the estimator; the second term is the *simple response variance*. We define the *index of inconsistency*

$$I = \frac{\sigma_\varepsilon^2}{\sigma_y^2 + \sigma_\varepsilon^2} \tag{2.8}$$

which measures the proportion of the total element variance which is due to response variability.

It is interesting to note that under this model, the index of inconsistency I is closely related to the correlation between successive observations on the same individual. Indeed

$$\underset{j}{Corr}\left(y_{j1},y_{j2}\right) = \frac{\sqrt{\underset{j}{E}\left(y_{j1},y_{j2}\right) - \underset{j}{E}\left(y_{j1}\right)\underset{j}{E}\left(y_{j2}\right)}}{\sqrt{\underset{j}{E}\left(y_{j1} - \bar{y}\right)^2 \underset{j}{E}\left(y_{j2} - \bar{y}\right)^2}}$$

$$= \frac{\sigma_y^2}{\sigma_y^2 + \sigma_\varepsilon^2} \qquad (2.9)$$

$$= 1 - I$$

In order to estimate σ_ε^2 we need to have at least two observations on at least a sample of individuals in the sample. This does not give any direct information on the values of the $\{\varepsilon_{ji}\}$. However the set of differences $\{y_{j1} - y_{j2}\}$ provide for us the values of $\{\varepsilon_{j1} - \varepsilon_{j2}\}$. The variance of $(\varepsilon_{j1} - \varepsilon_{j2})$ can be estimated simply and is

$$\sigma_{\varepsilon_{1,2}}^2 = \sigma_{\varepsilon_1}^2 + \sigma_{\varepsilon_2}^2 - 2\sigma_{\varepsilon_1\varepsilon_2}$$

If we assume, not unreasonably, that $\sigma_{\varepsilon_1}^2 = \sigma_{\varepsilon_2}^2 = \sigma_\varepsilon^2$, this gives

$$\sigma_{\varepsilon_{1,2}}^2 = 2\sigma_\varepsilon^2\left(1 - \rho_{\varepsilon_1\varepsilon_2}\right) \qquad (2.10)$$

We estimate σ_ε^2 by

$$\hat{\sigma}_\varepsilon^2 = \tfrac{1}{2}\hat{\sigma}_{\varepsilon_{1,2}}^2$$

The critical problem with this estimator is that there may be a correlation (usually positive in practice) between the response errors of the same individual on the two occasions; the respondent may for example remember some of the responses from the first interview, and tend to report the same answers in the re-interview. If the correlation is positive, $\hat{\sigma}_\varepsilon^2$ underestimates the simple response variance in the survey by a factor of $\left(1 - \rho_{\varepsilon_1\varepsilon_2}\right)$.

Data may be used to investigate whether such a positive correlation is present by comparing the variance of the response deviations for different time intervals between the interviews. This issue is considered in chapters 7 and 8.

Case II Correlated Interviewer Variance

The assumptions in case I are unrealistic in so far as the variable response errors $\{\varepsilon_{ji(s)}\}$ are assumed to be independent of one another. There are various factors which make this assumption unlikely to be appropriate. In particular, each interviewer carries out a number of interviews and it may be expected that the responses obtained from each individual in the interviewer's workload may be influenced in a similar way by the interviewer, and that, in consequence, the response errors for these individuals may be correlated. The model can be specified in such a way that this can be taken into account. Each observation is now denoted by $y_{iji(s)}$ where i denotes the interviewer. The assumptions are given below:

$$\begin{aligned} &\beta_{iji(s)} = \beta && \text{all } i, j \\ &V_\eta(\varepsilon_{iji(s)}) = \sigma_\varepsilon^2 && \text{all } i, j \\ &\mathrm{Cov}_\eta(\varepsilon_{iji(s)}, \varepsilon_{i'j'i(s)}) = && \rho_1 \text{ if } i = i' \\ &&& \rho_2 \text{ if } i \neq i' \end{aligned} \qquad (2.11)$$

The total variance of $\bar{y}_{..}$ under this model is

$$V_{p\eta}(\bar{y}_{..}) = \frac{\sigma_y^2}{n} + \frac{\sigma_\varepsilon^2}{n}\left[1 + \rho_1(m-1) + \rho_2 m(k-1)\right]$$

where m is the size of each interviewer's workload.

In order to estimate the correlated response variance due to the interviewers the survey design must be modified. The modification consists of applying interpenetration in at least part of the sample, as described in section 2.3.1. The basic principles can most easily be illustrated in the case of a simple random sample. A simple random sample s of size $n = km$ is selected from

the population; the sample is partitioned into k equal subsamples of size m - $s_1, s_2, \ldots, s_k$. Each subsample is allocated to a single interviewer. The label *(i,j)* is used to indicate that individual j belongs to the workload of interviewer i.

The usual estimator of the variance of the sample mean is

$$\hat{V}(\bar{y}_{..t(s)}) = (1-f)\frac{1}{n(n-1)} \sum_{i=1}^{k} \sum_{j \in s_i} (y_{itj(s)} - \bar{y}_{..t(s)})^2$$

It can be shown that the expected value of this estimator is

$$E_p E_\eta(\hat{V}(\bar{y}_{..t(s)}) = (1-f)\left[\frac{\sigma_y^2}{n}\frac{N}{N-1} + \frac{1}{n(n-1)}\sigma_\varepsilon^2\{(n-1) - \rho_1(m-1) - \rho_2(m)(k-1)\}\right]$$

If the sampling fraction is negligible, the bias of this estimator is given by

$$E_p E_\eta \hat{V}(\bar{y}_{..t(s)}) - V(\bar{y}_{..t(s)}) = \frac{1}{n-1}\left[-\rho_1(m-1)\sigma_\varepsilon^2 - \rho_2 m(k-1)\sigma_\varepsilon^2\right]$$

From the data we can calculate two linearly independent sums of squares:

1 the between-interviewers sum of squares, and

2 the within-interviewer sum of squares.

If we denote the mean between-interviewers sum of squares by C and the mean within-interviewer sum of squares by F, we can show that

$$E_p E_\eta[C] = \sigma_y^2 + \sigma_\varepsilon^2 [1+\rho_1(m-1) - \rho_2 m]$$

and (2.13)

$$E_p E_\eta[F] = \sigma_y^2 + \sigma_\varepsilon^2 [1-\rho_1]$$

Hence $\frac{1}{m}[C-F]$ provides a possible estimator of $\rho_1\sigma_\varepsilon^2$. In fact, under this model,

$$E\left[\frac{1}{m}[C-F]\right] = \sigma_\varepsilon^2[\rho_1 - \rho_2]$$

It is usually recommended as an estimator of $\rho_1\sigma_\varepsilon^2$ since ρ_2 can generally be assumed to be small. See, for example, Hansen, Hurwitz and Bershad (1961), Fellegi (1964) and Kish (1962).

It is worth noting that an almost unbiased estimator of the total simple variance $\left[\left(\sigma_y^2 + \sigma_\varepsilon^2\right)\right]$ is given by $\frac{1}{m}(C-F)+F$. In fact,

$$E_p E_\eta\left[\frac{1}{m}(C-F) + F\right] = \sigma_y^2 + \sigma_\varepsilon^2(1-\rho_2) .$$

In cases where σ_y^2 and σ_ε^2 cannot be disentangled (for example in the cases studies described in chapter 3) we use ρ_i (where the *i* denotes "interviewer") to represent $\rho_1\sigma_\varepsilon^2$.

Case III Simple and Correlated Response Variance

It is clear from cases I and II above that both re-enumeration and randomized allocation of respondents are necessary if both the simple and correlated response variances are to be estimated. A more elaborate survey design is necessary in this case. A general design is described by Fellegi (1964); the WFS examples in chapters 4, 5, and 6 used a variant of this design.

1 A *simple random sample* of $n = km$ units denoted by s, is selected *without replacement* from a population of N units.

2 The sample is partitioned, again at *random*, into k subsamples of m units each, denoted by $s_1, s_2, \ldots, s_k$.

3 Each subset is paired at *random* with another (different) subset so that if (s_1, s_{q1}),

$(s_2,s_{q\,2}),\ldots,(s_k,s_{q\,k})$ are the pairs then $q_1,\ldots,q_k$ exhaust the integers $1,\ldots,k$.

4 There are k interviewers, numbered from 1 to k and the k pairs of subsets are allocated at *random* to these. Denote by $(s_{i(1)},s_{i(2)})$ the pair allocated to the i-th interviewer; $s_{i(1)}$ and $s_{i(2)}$ are respectively the first and second assignments of the i-th interviewer.

Steps (1), (2) and (3) define the experimental design. The randomization is artificial and the probability distribution associated with it can, in principle, be calculated without too much difficulty. Let Q denote the outcome obtained from steps (1), (2), (3) and (4). Let $p(Q)$ denote the probability of obtaining the outcome Q.

5 Each interviewer interviews his first assignment and this constitutes the original survey.

6 Each interviewer interviews his second assignment and this constitutes the repeat survey.

Steps (5) and (6) define the measurement process. The η-distribution is introduced here.

Let $y_{ijt(Q)}$ be the observed value of unit j at trial t given that the outcome Q was obtained. If, given Q, $j \notin s_{i(t)}$ $(t=1,2,)$, $y_{ijt(Q)}$ need not be defined at all. The model is analogous to the model (2.11). Thus we write

$$y_{ijt(Q)} = y_j + \beta_{ijt(Q)} + \varepsilon_{ijt(Q)} \tag{2.14}$$

The implications of the model (2.14) depend on the specification of the $\{\beta_{ijt(Q)}\}$ and the probability distribution of the $\{\varepsilon_{ijt(Q)}\}$.

Once again we assume that the $\beta_{ijt(Q)}$ are constant. The probability distribution of the $\{\varepsilon_{ijt(Q)}\}$ must also be specified. We make two simplifying assumptions:

$$\begin{aligned} E_\eta\left(\varepsilon_{ijt(Q)}\right) &= 0 \\ V_\eta\left(\varepsilon_{ijt(Q)}\right) &= \sigma^2_{\varepsilon_t} \end{aligned} \tag{2.15}$$

We must also define all the other second-order moments of the $\{\varepsilon_{ijt(Q)}\}$. In general, using (2.14)

$$Cov_{\eta}\left(\varepsilon_{ijt(Q)}\varepsilon_{i'j't'(Q)}\right) = \rho_{ijt,i'j't'}\sigma_{\varepsilon_j}\sigma_{\varepsilon_{j'}}$$

There are many ways in which the $\{\rho_{ijt,i'j't'}\}$ could be specified. The survey design however permits us to define easily the following correlation coefficients. Let i^o be the interviewer who carries out the i-th interviewer's first workload in the re-interview survey. Then, $s_{i(1)} = {}^{\square}{}_{i^o(2)}$.

$$\rho_{ijt,i'j't'(Q)} = \begin{cases} \rho_{1t} & i'=i, j'\neq j \ \varepsilon \ s_{i(t)} \\ & \text{(same interviewer, different units, same trial)} \\ \rho_{2t} & i'\neq i, j \ \varepsilon \ s_{i(t)},\ j' \ \varepsilon \ s_{i'(t)} \\ & \text{(different interviewers, different units, same trial)} \\ \rho_3 & i'=i^o, j'=j,\ t=1\ t'=2 \\ & \text{(same unit, different interviewers, different trials)} \\ \rho_4 & i'=i, j \ \varepsilon \ s_{i(1)}, j' \ \varepsilon_s s_{i'(2)}\ t=1,\ t'=2 \\ & \text{(same interviewer, different units, different trials)} \\ \rho_5 & i'=i^o, j'\neq j \ \varepsilon \ s_{i(1)},\ t=1,\ t'=2 \\ & \text{(same subsample, different units, different trials)} \\ \rho_6 & i'\neq i, i^o, j \ \varepsilon \ s_{i(1)}, j \ \varepsilon \ s_{i'(2)},\ t=1,\ t'=2 \\ & \text{(different interviewers, different units, different subsamples, different trials).} \end{cases} \tag{2.16}$$

Note that ρ_{1t} and ρ_{2t} correspond to ρ_1 and ρ_2 in case II; and ρ_3 is equivalent to $\rho_{\varepsilon 1,\varepsilon 2}$ in case I. There are eight correlation coefficients defined in (2.16). In addition to these we also wish to estimate $\sigma^2_{\varepsilon 1}$, $\sigma^2_{\varepsilon 2}$ and σ^2_y . Thus we have eleven parameters in all. From the data we can obtain seven linearly independent sums of squares. These are

1 For each of the two surveys

(a) Between interviewers

(b) Within interviewers

2 Between the two surveys

(a) within sampling units

(b) within subsamples, between interviewers

(c) within interviewers, between subsamples (this sum of squares is linearly independent of the previous ones if and only if $k>2$).

The corresponding mean squares are listed as follows:

$$C_t = \frac{m}{k-1}\sum_{i=1}^{k}(\bar{y}_{i.t(Q)} - \bar{y}_{..t(Q)})^2,\ t = 1,2$$

$$F_t = \frac{1}{k(m-1)}\sum_{i=1}^{k}\sum_{j\varepsilon s_{i(t)}}(y_{itj(Q)} - y_{i.t(Q)})^2,\ t = 1,2$$

$$L = \frac{1}{2(km-1)}\sum_{i=1}^{k}\sum_{j\varepsilon s_{i(t)}}(y_{itj(Q)} - y_{i\circ j2(Q)} - \bar{y}_{..1(Q)} + \bar{y}_{..2(Q)})^2$$

$$M = \frac{m}{2(k-1)}\sum_{i=1}^{k}(\bar{y}_{i.1(Q)} - \bar{y}_{i\circ.2(Q)} - \bar{y}_{..1(Q)} + \bar{y}_{..2(Q)})^2$$

$$P = \frac{m}{2(k-1)}\sum_{i=1}^{k}(\bar{y}_{i.1(Q)} - \bar{y}_{i.2(Q)} - \bar{y}_{..1(Q)} + \bar{y}_{..2(Q)})^2$$

If $k = 2$ the seven mean squares above are not linearly independent. In fact, in this case, $M+P=C_1+C_2$. Fellegi (1964) suggests that, instead of using M, it may be more convenient to use

$$R = \frac{1}{2k(m-1)}\sum_{i=1}^{k}\sum_{j\varepsilon s_{i(1)}}(y_{ij1} - y_{i\circ j2} - \bar{y}_{i.1} + \bar{y}_{i\circ.2})^2 = \frac{km-1}{k(m-1)}L - \frac{k-1}{k(m-1)}M$$

The design has provided us with seven linearly independent sums of squares to estimate eleven parameters. It is not therefore possible to obtain unbiased estimators for all of the parameters. However, for the surveys with which we are concerned here, some further simplifying assumptions can be made:

$$\sigma^2_{\varepsilon_1} = \sigma^2_{\varepsilon_2} = \sigma^2_{\varepsilon} \qquad \rho_{11} = \rho_{12} = \rho_1 \qquad \rho_{21} = \rho_{22} = \rho_2 \tag{2.17}$$

These assumptions correspond to an assertion that the essential survey conditions over the two surveys are the same. If (2.17) holds, then the expected values of the mean squares listed above are:

$$E_pE_\eta(C_1) = E_pE_\eta(C_2) = \sigma_y^2 + \sigma_\varepsilon^2\left[1 + (m-1)\rho_1 - m(k-1)\rho_2\right]$$

$$E_pE_\eta(F_1) = E_pE_\eta(F_2) = \sigma_y^2 + \sigma_\varepsilon^2\left[1 - \rho_1\right]$$

$$E_pE_\eta(L) = \sigma_\varepsilon^2\left[1 - \frac{m-1}{km-1}\rho_1 - \frac{m(k-1)}{km-1}\rho_2 - \rho_3 + \frac{m}{km-1}\rho_4 + \frac{m-1}{km-1}\rho_5 + \frac{m(k-2)}{km-1}\rho_6\right]$$

$$E_pE_\eta(M) = \sigma_\varepsilon^2\left[1 + (m-1)\rho_1 - m\rho_2 - \rho_3 + \frac{m}{k-1}\rho_4 - \frac{m-1}{k-1}\rho_5 + \frac{m(k-2)}{k-1}\rho_6\right]$$

$$E_pE_\eta(P) = \frac{k}{k-1}\sigma_y^2 + \sigma_\varepsilon^2\left[1 + (m01)\rho_1 - m\rho_2 + \frac{1}{k-1}\rho_3 - m\rho_4 + \frac{m-1}{k-1}\rho_5 + \frac{m(k-2)}{k-1}\rho_6\right.$$

$$E_pE_\eta(R) = \sigma_\varepsilon^2\left[1 - \rho_1 - \rho_3 + \rho_5\right]$$

In our analysis we use

$$C = \tfrac{1}{2}(C_1 + C_2) \text{ and } F = \tfrac{1}{2}(F_1 + F_2).$$

We now have five linearly independent sums of squares to estimate eight parameters. The system of equations we have to solve may be written as:

$$E_pE_\eta\begin{bmatrix}C\\F\\L\\R\\P\end{bmatrix} = \begin{bmatrix}1 & 1 & m-1 & -m & 0 & 0 & 0 & 0\\ 1 & 1 & -1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & -\frac{m-1}{km-1} & -\frac{m(k-1)}{km-1} & -1 & \frac{m}{km-1} & \frac{m-1}{km-1} & \frac{m(k-2)}{km-1}\\ 0 & 1 & -1 & 0 & -1 & 0 & 1 & 0\\ \frac{k}{k-1} & 1 & m-1 & -m & \frac{1}{k-1} & -m & \frac{m-1}{k-1} & \frac{m(k-2)}{k-1}\end{bmatrix}\begin{bmatrix}\sigma_y^2\\ \sigma_\varepsilon^2\\ \rho_1\sigma_\varepsilon^2\\ \rho_2\sigma_\varepsilon^2\\ \rho_3\sigma_\varepsilon^2\\ \rho_4\sigma_\varepsilon^2\\ \rho_5\sigma_\varepsilon^2\\ \rho_6\sigma_\varepsilon^2\end{bmatrix}$$

It should be possible to find estimators of five of the eight parameters so that the biases in the estimators are in terms of the remaining three parameters. The most important parameters

$$[T] = \begin{bmatrix} \frac{1}{m} & -\frac{1}{m} & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ -\frac{1}{m}\left(\frac{k-1}{k-2}\right) & \frac{1}{m} & \frac{km-1}{km}\left(\frac{k-1}{k-2}\right) & -\frac{1}{m}\frac{km-m-1}{k-2} & \frac{1}{km}\left(\frac{k-1}{k-2}\right) \\ \frac{1}{m} & -\frac{1}{m} & 0 & 1 & 0 \\ 0 & -\frac{1}{m} & -\frac{km-1}{km} & 1 & \frac{k-1}{km} \end{bmatrix}$$

Consequently the following five linearly independent combinations of parameters have unique unbiased estimators:

1	$(\rho_1 - \rho_2)\sigma_\varepsilon^2$
2	$(\sigma_y^2 + \rho_3\sigma_\varepsilon^2 - \rho_5\sigma_\varepsilon^2)$
3	$-\rho_5\sigma_\varepsilon^2 + \rho_6\sigma_\varepsilon^2$
4	$\sigma_\varepsilon^2(1 - \rho_2 - \rho_3 + \rho_5)$
5	$-\rho_4\sigma_\varepsilon^2 + \rho_5\sigma_\varepsilon^2$

The estimators for 1 to 5 are:

$$E_1 = \frac{1}{m}[C - F]$$

$$E_2 = F - R$$

$$E_3 = -\frac{1}{m}\left(\frac{k-1}{k-2}\right)C + \frac{1}{m}F + \frac{km-1}{km}\left(\frac{k-1}{k-2}\right)L - \frac{1}{m}\left(\frac{km-m-1}{k-2}\right)R + \frac{k-1}{km(k-2}P$$

$$E_4 = \frac{1}{m}[C-F] + R$$

$$E_5 = -\frac{1}{m}f - \frac{km-1}{km}L + R + \frac{k-1}{km}P$$

On the assumption that ρ_2 and ρ_5 are negligibly small, and disregarding ρ_3 for the moment, we obtain the following estimators for the principal parameters of the system:

$$\hat{\sigma}_y^2 = F - R \tag{2.18}$$

$$\hat{\sigma}_\varepsilon^2 = R + \frac{1}{M}[C - F] \tag{2.19}$$

$$\hat{\rho}_1\hat{\sigma}_\varepsilon^2 = \frac{1}{m}[C - F] \tag{2.20}$$

$$\hat{\rho}_4\hat{\sigma}_\varepsilon^2 = \frac{1}{m}F + \frac{km-1}{km}L + R + \frac{k-1}{km}P \tag{2.21}$$

$$\hat{\sigma}_y^2 + \hat{\sigma}_\varepsilon^2 = \frac{1}{m}[C - F] + F \tag{2.22}$$

and hence,

$$\hat{\rho}_1 = \frac{\frac{1}{m}[C - F]}{R + \frac{1}{m}[C - F]} \tag{2.23}$$

$$\hat{I} = \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2 + \hat{\sigma}_y^2} = \frac{R + \frac{1}{m}[C - F]}{\frac{1}{m}[C - F] + F} \tag{2.24}$$

$$\text{and} \quad \hat{\rho}_1 I = \frac{\frac{1}{m}[C - F]}{\frac{1}{m}[C - F] + F} \tag{2.25}$$

Apart from the problem of ρ_3, the biases in the estimators derived from the data include only terms in ρ_2 and ρ_5, which may reasonably be assumed to be small. However in social surveys ρ_3 is not generally negligible, since it is a measure of the recall effect, and it is usually positive.

Hence, ignoring terms in ρ_2 and ρ_5,the estimate of σ_y^2 has an expected value

$$E\left(\hat{\sigma}_y^2\right) = \sigma_y^2 + \rho_3\sigma_\varepsilon^2$$

Similarly, the estimate of σ_ε^2 has an expected value

$$E\left(\hat{\sigma}2_{\varepsilon}^{2}\right) = \sigma_{\varepsilon}^{2} - \rho_3\sigma_{\varepsilon}^{2}$$

Thus $\hat{\sigma}_y^2$ overestimates σ_y^2 and $\hat{\sigma}_\varepsilon^2$ underestimates σ_ε^2 . However the total simple variance $\left(\sigma_y^2 + \sigma_\varepsilon^2\right)$ can be estimated with negligible bias using $\frac{1}{m}[C - F] + F$.

It is only by introducing information external to the model presented here - for example, the time interval between two interviews - that any assessment of ρ_3 can be obtained. Two approaches to this issue are discussed later: analysis by time interval in chapter 7, and a modelling perspective in chapters 7 and 8.

Case IV Interaction between Sampling and Response Deviations

It may be postulated that an interviewer's workload will influence the response deviations obtained by the interviewer. If this is so, then the sampling deviations $\{\Delta_j\}$ may be correlated with the response deviations (the $\{\varepsilon_{ijt}\}$). The structure of the model will thus be more complex and the expected values of the seven mean squares will be as follows:

$$_{7\eta}\begin{bmatrix} C \\ F \\ L \\ R \\ P \end{bmatrix} = \begin{bmatrix} 1 & 1 & (m-1) & (m-1) & -m & 0 & 0 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -\frac{m}{km-1} & -\frac{m(k-1)}{km-1} & -1 & \frac{m}{km-1} & \frac{m-1}{km-1} & \frac{m(k-2)}{km-1} \\ 0 & 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ \frac{k}{k-1} & 1 & \frac{k(m-1)}{k-1} & m-1 & -m & \frac{1}{k-1} & -m & \frac{m-1}{k-1} & \frac{m(k-2}{k-1} \end{bmatrix} \begin{bmatrix} \sigma_y^2 \\ \sigma_\varepsilon^2 \\ 2\alpha\sigma_y\sigma \\ \rho_1\sigma_\varepsilon^2 \\ \rho_2\sigma_\varepsilon^2 \\ \rho_3\sigma_\varepsilon^2 \\ \rho_4\sigma_\varepsilon^2 \\ \rho_5\sigma_\varepsilon^2 \\ \rho_6\sigma_\varepsilon^2 \end{bmatrix}$$

where $\alpha\sigma_y\sigma_\varepsilon = E_{p\eta}[\Delta_j\varepsilon_{ijt}]$

The set of transformations previously applied to the coefficients matrix in case III now gives:

$$[T]\ E_{pn}\begin{bmatrix} C \\ F \\ L \\ R \\ P \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 1 & 1 & 0 & -1 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \sigma_y^2 \\ \sigma_\varepsilon^2 \\ 2\alpha\sigma_y\sigma_\varepsilon \\ \rho_1\sigma_\varepsilon^2 \\ \rho_2\sigma_\varepsilon^2 \\ \rho_3\sigma_\varepsilon^2 \\ \rho_4\sigma_\varepsilon^2 \\ \rho_5\sigma_\varepsilon^2 \\ \rho_6\sigma_\varepsilon^2 \end{bmatrix}$$

Hence the estimators E_1 to E_5 in case III now have the following expectations:

$$\begin{aligned} E(E_1) &= \rho_1\sigma_\varepsilon^2 + 2\alpha\sigma_y\sigma_\varepsilon - \rho_2\sigma_\varepsilon^2 \\ E(E_2) &= \left[\sigma_y^2 + \rho_3\sigma_\varepsilon^2\right] - 2\alpha\sigma_y\sigma_\varepsilon \\ E(E_3) &= -2\alpha\sigma_y\sigma_\varepsilon - \rho_5\sigma_\varepsilon^2 + \rho_6\sigma_\varepsilon^2 \\ E(E_4) &= \left[\sigma_\varepsilon^2 - \rho_3\sigma_\varepsilon^2\right] + 2\alpha\sigma_y\sigma_\varepsilon - \rho_2\sigma_\varepsilon^2 + \rho_5\sigma_\varepsilon^2 \\ E(E_5) &= 2\alpha\sigma_y\sigma_\varepsilon - \rho_4\sigma_\varepsilon^2 + \rho_5\sigma_\varepsilon^2 \end{aligned}$$

Thus we can construct estimators of the most important parameters in the model with biases which involve only ρ_2, ρ_5 and ρ_6, all of which may reasonably be assumed to be negligible.

Parameter	Estimator	Bias
$\sigma_y^2 + \sigma_\varepsilon^2$	$E_2 + E_4$	$-\rho_2\sigma_\varepsilon^2$
$\sigma_y^2 + \rho_3\sigma_\varepsilon^2$	$E_2 - E_3$	$-\rho_6\sigma_\varepsilon^2$
$\sigma_\varepsilon^2 - \rho_3\sigma_\varepsilon^2$	$E_4 + E_3$	$-\rho_2\sigma_\varepsilon^2 + \rho_6\sigma_\varepsilon^2$

$2\alpha\sigma_y\sigma_\varepsilon$	$-E_3$	$\rho_5\sigma_\varepsilon^2 - \rho_6\sigma_\varepsilon^2$
$\rho_1\sigma_\varepsilon^2$	$E_1 + E_3$	$-\rho_2\sigma_\varepsilon^2$
$\rho_4\sigma_\varepsilon^2$	$-[E_5 + E_3]$	$-\rho_6\sigma_\varepsilon^2$

The problem still remains with the estimation of σ_y^2 and σ_ε^2 since there is no way of separating the impact of ρ_3. However, $\sigma_y^2 + \rho_3\sigma_\varepsilon^2$ and $\sigma_\varepsilon^2 - \rho_3\sigma_\varepsilon^2$ can be estimated with negligible bias.

2.3.2 Categorical and Ordinal Data

The proportion of the sample in any single category for a categorical or ordinal variable can be treated in terms of the general model described in section 2.3.1. The relative simplicity of ordinal and categorical data, however, also provides an opportunity both to use simpler measures and to obtain simple forms of some of the measures previously described.

For a categorical variable the responses obtained from the two interviews may be represented by the square matrix $\{p_{ij}\}$ where p_{ij} is the proportion of the elements classified in category i according to the first interview and in category j according to the second interview. The diagonal of this square matrix, with entries p_{ii}, contains the cases of exact agreement. The simplest measure of reliability (bivariate agreement) is the *index of crude agreement*

$$A = \Sigma p_{ii}. \tag{2.26}$$

which is the proportion of the cases classified identically by the two observations. This index has considerable descriptive value. In later tables (chapter 6 section 6.2) we present A and its complement, the index of crude disagreement

$$D = 1 - A \tag{2.27}$$

This crude index has a fairly serious drawback, however: it does not take into account the fact that some agreement will occur by chance even if the measurement is completely unreliable (random). The extent of chance agreement depends upon the two marginal distributions $\{p_{i.} (= \Sigma_j p_{ij})\}$ and $\{p_{.j} (= \Sigma_i p_{ij})\}$. One approach, due to Cohen (1960), is to define an index of consistency, kappa, of the form

$$\begin{aligned} \kappa &= 1 - \frac{\textit{observed disagreement}}{\textit{expected disagreement}} \\ &= 1 - \frac{1 - p_o}{1 - p_e} = \frac{p_o - p_e}{1 - p_e} \end{aligned} \tag{2.28}$$

Under the baseline constraint of independence between the two observations, we have

$$p_e = \left(\sum_i p_{ii} \right)_e = \sum p_{i.}\, p_{.i}$$

giving

$$\kappa = \sum_i (p_{ii} - p_{i.}\, p_{.i}) \Big/ \left(1 - \sum_i p_{i.}\, p_{.i} \right) \tag{2.29}$$

While (2.29) is a more appropriate measure of reliability, particularly in the presence of skewness in the distribution across categories, it can be misleading in situations where a single category dominates the marginal distributions: the value of κ will in this case tend to suggest a low level of consistency if any elements occur off the diagonal. Another point to note in relation to (2.29) is that it would be inappropriate to use κ on its own to describe the level of agreement since it conditions on the observed marginals. The degree of agreement between the marginals is in itself an important component of the observation process. One of a number of possible measures of the disagreement between marginal distributions themselves is

$$B = \frac{2}{\pi} \cos^{-1} \left[\sum_i (p_{i.}\, p_{.i})^{1/2} \right] \tag{2.30}$$

with value `1' indicating complete disagreement, and `0' complete agreement, between the two marginal distributions.

The measures (2.26) - (2.29) above are the traditional measures used in cases of multiple

observation of the same individual. It is possible also to define simply two other measures.

The index of inconsistency (I), already defined in (2.8), has a particularly simple form for a binary variable. The results of the two interviews can be summarized as follows:

First interview	Re-interview		
	Yes	No	Total
Yes	a	b	$a + b = p_1$
No	c	d	$c + d = q_1$
Total	$a + c = p_2$	$b + d = q_2$	1

It follows that:

$$E(p_1q_1) = E(p_2q_2) = \sigma_y^2 + \sigma_\varepsilon^2$$
$$E(b+c) = 2\sigma_\varepsilon^2(1 - \rho_{\varepsilon_1\varepsilon_2})$$

Given that $\rho_{\varepsilon_1\varepsilon_2}$ cannot be estimated, a consistent estimator of I is provided by

$$\hat{I} = \frac{b + c}{p_1q_1 + p_2q_2} \qquad (2.31)$$

An alternative measure, which takes into account the magnitude of the proportion p, is given by

$$R = \frac{b + c}{\frac{1}{2}(p_1 + p_2)^2} \qquad (2.32)$$

which is a measure of the absolute increase of the relvariance of p due to the simple response variance.

The measures (2.26) - (2.32) described above apply to any level of measurement of the classification variable: categorical (nominal), ordered or metric. When the scales are categorical, any deviation from the diagonal constitutes disagreement. When the scales are ordinal, interval or ratio, any measure of agreement should take into account the *degree* of disagreement, which is a function of the difference between scale values. We can modify

(2.26) by redefining `agreement' to mean that the two interviews obtain values within some acceptable distance (k units) of each other

$$A_k = \sum_{|i-j| \le k} p_{ij} = 1 - D_k \tag{2.33}$$

Cohen (1968) introduced a modified form of κ which allows for scaled disagreement or partial credit in terms of weights w_{ij} which reflect the contribution of each cell in the table to the degree of disagreement:

$$\kappa_w = \frac{p_o^* - p_e^*}{1 - p_e^2} \tag{2.34}$$

where

$$p_o^* = \sum_{i,j} w_{ij} p_{ij} \; ; \; p_e^* = \sum_{i,j} (w_{ij} p_{i.} p_{.j})$$

Any monotonically decreasing function of the differences between the scale values of i and j can be used as weights.

Cicchetti (1972, 1973) suggests the use of the following weights:

$$\text{for ordinal data, } w_{ij} = 1 - |i-j|/(L-1) \tag{2.35}$$

where L is the number of categories, and

$$\text{for metric data, } w_{ij} = 1 - (i-j)^2 / (L-1)^2 \tag{2.36}$$

Under observed marginal symmetry, κ_w with weights (2.36) is precisely equal to the product-moment correlation coefficient for the integer-valued categories. Furthermore, under the assumption of the random effects model, the estimate of the intra-class correlation coefficient is asymptotically equal to κ_w (Cohen (1968); Fleiss and Cohen (1973)). These measures are discussed in more detail in Landis and Koch (1976).

Chapter 3 THREE CASE STUDIES IN INTERVIEWER EFFECT

Introduction

In this chapter three case studies in the estimation of interviewer variance are presented, each of which provides a basis for generalization in later chapters. All three are examples of *interpenetration without reenumeration*; this is the second approach described in section 2.3.1. In all cases therefore the only estimable quantities are $\{\rho_1\sigma^2_\epsilon\}$ and $\{\sigma^2_y + \sigma^2_\epsilon\}$. The ratio of these quantities is the usual estimate of the intra-interviewer correlation coefficient ρ_i (the details of the estimation and the model are presented in Case II of section 2.3.2). The studies cover, in subject matter, job satisfaction and work-related variables in Irish manufacturing industry, general health questions and attitudes towards noise in a community survey, and a wide variety of socio-economic variables in a national British survey. The populations are respectively workers in Irish manufacturing industry, residents of neighbourhoods near Heathrow airport (London), and adults in great Britain. The studies were carried out in 1969-70, 1978, and 1992-3. More comprehensive reports on the studies and the findings are available in O'Muircheartaigh (1976), O'Muircheartaigh and Wiggins (1981), and O'Muircheartaigh and Campanelli (1998).

3.1 The Absenteeism Study

This interviewer variance study is part of a national survey of absenteeism and labour turnover in Ireland. The substantive results are described in O'Muircheartaigh (1974). The survey was undertaken in two phases. In the first phase a probability sample of 180 industrial establishments was selected and basic data on absenteeism and labour turnover were collected from each. Using this information, a subsample of establishments in areas of particular interest was selected and examined more intensively by means of interviews with a sample of managers, supervisors and workers. As preparation for this second phase, a study of absenteeism was carried out in one large Irish industrial concern. Besides serving as a pilot study for the second phase survey, this study provided data for an examination of rates of absence within the company chosen, and an investigation of

response errors was incorporated into the study design. The interviewer variance design was repeated in the main (second) phase of the survey.

3.1.1 The Data

Two sets of data are used in this analysis. The first set was obtained in 131 interviews with male production workers in a large industrial organization in Dublin. This was the final pilot study in the design of the phase II questionnaire. Five interviewers were used in the study and the workers were allocated randomly among the interviewers. Interviews were arranged for the interviewers by the departmental managers, and once the worker had consented to come for interview, no refusals were encountered. The interviews took place during working hours in offices provided by the company. The questionnaire - which closely resembled the phase II questionnaire - dealt with various aspects of the workers' satisfaction with their jobs and with the company. In order to avoid coding variability, all the questionnaires were coded by one experienced coder.

The second set of data consists of the responses to the interviews in phase II. Seven hundred and sixty seven interviews were carried out in this phase. In the cases where it was possible to do so, two interviewers were allocated to each firm and the sample members were allocated randomly between them. In four firms, however, this was not possible and these firms had to be excluded from this analysis. The thirteen remaining firms may be divided into two classes: (i) two interviewers were allocated to each of two firms in five cases (i.e. ten firms altogether) and shared the interviewing in each firm; (ii) in the three remaining firms, two interviewers were allocated to each firm and worked only in that firm. Thus, sixteen interviewers in all were involved in this data set.

The questionnaire items are examined one by one in terms of the interviewer effect. This part of the analysis has two aims: (i) to provide a general estimate of the magnitude of the interviewer variance; (ii) to identify individual items or interviewers which are particularly susceptible to error.

3.1.2 Empirical results

The pilot study

Previous research (Gales and Kendall, 1957, for instance) had shown that it is useful to distinguish "factual" from "opinion" items. There were five factual items in the questionnaire. Table 3.1 below gives the value of $\hat{\rho}_i$ for each item. Recall that ρ_i is the intra-interviewer correlation coefficient (corresponding to $\rho_I\sigma_\varepsilon^2$ in the fully specified model - case II or case III in section 2.3.2).

Table 3.1: Interviewer effect estimates for factual items in the absenteeism pilot survey

Item	$\hat{\rho}_i$
Age	-0.02
Marital status	+0.00
Length of service	-0.02
No. of children	+0.01
Skill level	+0.20

For the number of degrees of freedom available in this study, values of $\hat{\rho}_i \geq 0.08$ are significant at the 10% level; $\hat{\rho}_i \geq 0.10$ at the 5% level, and $\hat{\rho}_i \geq 0.16$ at the 1% level. The fifth item in this group was considered to be factual - the workers were graded as skilled/semi-skilled/unskilled by the interviewers at the end of the interview. The very high value of $\hat{\rho}_i$ and its low level of probability if no interviewer effect were present (prob < 0.001) made further investigation imperative. To examine the structure of the interviewer effect for this question, the results were cross-tabulated by interviewer. From this analysis is was discovered that one interviewer classified only two workers as unskilled (6%) whereas the lowest proportion classified as unskilled by any of the other interviewers was 36%. By checking the job descriptions it became apparent that the instructions to the interviewers on this question had not been clearly understood. The coefficients for the factual items illustrate the point made

by Gray (1956) that interviewer variability found on factual questions suggests that the interviewer briefing is at fault. There is no evidence from the other four items that any interviewer effect is present.

Two recall questions, each relating to three reference periods, were included in the questionnaire. The items and the corresponding $\hat{\rho}_i$ values are presented in table 3.2.

Table 3.2 Interviewer variance estimates for recall items in the absenteeism pilot survey

Item			$\hat{\rho}_i$
1.	On how many occasions were you absent		
	(i)	in the last 6 months	-0.04
	(ii)	in the last 3 months	-0.02
	(iii)	in the last 1 month	+0.03
2.	How many days in all were you absent		
	(i)	in the last 6 months	-0.02
	(ii)	in the last 3 months	-0.02
	(iii)	in the last 1 month	+0.01

The low values for these questions are reassuring. The questions dealt with a sensitive area of organisational behaviour and were important to the study since the main emphasis of the study was on absenteeism. As four of the six items have negative $\hat{\rho}_i$'s and the absolute sizes of the $\hat{\rho}_i$'s are small there is no evidence that interviewer variability contributes to the overall variance on these items.

There were 102 attitudinal items used in the study. Table 3.3 below gives the distribution of the $\hat{\rho}_i$-values for these items. The range of $\hat{\rho}_i$ for the 102 items was 0 to 0.30.

Table 3.3: Interviewer variance estimates for attitudinal items in the absenteeism pilot survey

$\hat{\rho}_i$	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10	0.15	0.30
No. of items	40	6	7	6	6	4	4	3	7	5	2	9	3

The items are similar in content to the attitudinal items in the first Kish study (Kish, 1962) in which, of the 46 items, 9, 15, and 17 were significant at the 1%, 5%, and 10% levels respectively. In this investigation, of the 102 attitudinal items, 3, 14, and 26 are significant at the 1%, 5%, and 10% levels. The effect of the interviewer variance is considerable. Using the sample $\hat{\rho}_i$ as the best estimate of the population ρ_i, the contribution of the interviewer effect to the overall variance is given by $\{1+\hat{\rho}(m\text{-}1)\}$. In this study $m = 26$ and a value of $\hat{\rho} = +0.04$ will have the effect of doubling the variance. In the case of the attitudinal items, 26 had a value greater than +0.08, and a $\hat{\rho}$ of +0.08 means a trebling of the variance or an increase of 70% in the standard error.

There is some evidence to suggest that different types of items are affected to a different extent by interviewers. Gales and Kendall (1957) found that ambiguity in a question leads to high variability. Hanson and Marks (1958) found that contributory factors were (i) interviewer "resistance" to a question, (ii) relatively high ambiguity, (iii) the extent to which additional questioning (probing) tends to alter initial respondent's replies. Kish (1962) examined three categories of items - critical, ambiguous and other - but found no clear pattern. In this investigation a comparison was made between closed and open attitude questions, there being 58 closed and 44 open items used in the study. Table 3.4 gives the cumulative distributions of the $\hat{\rho}_i$-values for the two sets of items.

Table 3.4 Cumulative percentages of ρ_i's for closed and open items

	0.00	0.01	0.05	0.06	0.07	0.09	0.15	0.30
Cum. percent of closed items	43	50	76	81	83	91	98	100
Cum. percent of open items	34	39	57	59	63	79	95	100

A test of the difference between the two distributions using the Kolmogorov-Smirnov one-tailed test gives a significance level of 0.10. A one-tailed test is appropriate in this case since open questions can be expected to be more sensitive to interviewer variability for two reasons. Firstly, there is no frame of reference for open questions and secondly, the interviewer selects a part of the respondent's reply to record. The result of the test is not conclusive. However, the open questions do show greater $\hat{\rho}$ values and if this pattern is supported by further evidence, the result is of some substantive importance.

The main survey

For the analysis of the items used in the main study, a different approach was necessary. A separate analysis of variance was carried out for each of the eight cases - the five pairs of firms and the three single firms - and values of $\hat{\rho}_i$ were obtained for each item for each case. There were nineteen `factual' items and seventy-seven attitudinal items on the questionnaire.

In the case of the factual items, the results for the main study were reassuring. All of the $\hat{\rho}_i$ values were small and none were significant. The form of the item on skill level and the briefing had been changed as a result of the analysis of the pilot study responses and this seems to have reduced the sensitivity of this item to interviewer effect.

In the case of the more complicated `factual' items in the questionnaire, however, all the results were not as satisfactory. For the item "number of people in work group" a value of $\hat{\rho}_i$ significantly greater than zero was obtained in three of the eight cases. This result serves to illustrate the difficulty of defining factual items carefully enough and conveying the definition to the interviewers. On the whole, however, the values of $\hat{\rho}_i$ for the factual items were small and did not show any significant interviewer effect.

For the attitudinal items in the main study the distribution of the values of $\hat{\rho}_i$ was similar to that for the pilot study, although the range of $\hat{\rho}_i$ was wider. This, however, can be

explained by the very small number of interviewers *(k=2)* used in each firm, and the resultant large variability in the estimates of ρ_i. It should be noted that the value of Wilks' lambda obtained to test the significance of each vector of attitudinal items was significant in twenty two of the twenty four cases [3 subsets of attitudinal items for each of the eight cases].

3.1.3 Summary of results and conclusions for the absenteeism study

In this investigation, the effect of interviewer variance was examined for various types of item. This effect was measured in terms of $\hat{\rho}_i$, the intra-interviewer correlation coefficient. The significance of the values of $\hat{\rho}_i$ was tested using the *F* test. For factual items the results agree with previous findings that, in general, no interviewer effect is present in these cases. In one case in the pilot study, a strong interviewer effect was evident but this was found to be due to inadequate briefing of the interviewers and was corrected for the main study. This provides further evidence that great care is needed in interviewer briefing even on factual items. For attitudinal items, there is strong evidence of the presence of a significant interviewer effect. The level of interviewer effect may be higher for open items than for closed items. The effect is significant both statistically and substantively. The range of $\hat{\rho}_i$ values was wider than the range of previously published studies. This is probably due primarily to the small number of interviewers used ($k=5$ and $k=2$).

The magnitude of interviewer variability also has an important implication for research design. With an interviewer effect ρ the variance of the sample mean for a sample size n can be expressed as

$$V(\bar{y}_n) = [1 + \rho(m-1)]V_o$$

where V_o is the variance of a sample of size n when no interviewer effect is present. For a value of $\rho = \frac{1}{m-1}$ we have $V(\bar{y}_n) = 2V_o$.

If we double the sample size, leaving the number of interviewers constant we have

$$V(\bar{y}_{2n}) = \left[1 + \frac{2m-1}{m-1}\right]\frac{V_o}{2} > \frac{3}{2} V_o$$

If we leave the sample size constant, and double the number of interviewers, we have

$$V'(\bar{y}_n) = \left[1 + \frac{\frac{m}{2} - 1}{m-1}\right] V_o < \frac{3}{2} V_o$$

Therefore, we can achieve approximately the same decrease in variance by doubling the number of interviewers as we can be doubling the sample size, although a deterioration in the quality of the interviewers may take place if the number of interviewers is expanded too far. In an ongoing survey operation with a permanent field-staff, the result implies that as many interviewers as possible should be in each study. Kish (1962) presents a model, from which the optimum size of interviewer workload can be estimated for different values of ρ, which takes into account the cost of training or briefing an interviewer for a particular study.

Finally, the investigation shows two ways in which a study of response errors can be used in a pilot study. Firstly, items with high interviewer effect can be examined to determine whether the structure of the item is at fault or whether one (or more) of the interviewers may have misinterpreted the instructions. Secondly, and equally important from a practical point of view, the study of interviewer variability and of other variable response errors may be used as a check to eliminate inefficient (or perhaps merely different) interviewers from the field force.

3.2 The Noise Annoyance Survey

Data collected from community surveys are generally obtained from structured interviews with the respondents. The data obtained and the conclusions reached depend on the quality of the data collection process. Numerous studies in mental health surveys have focused on respondent role behaviour and, in particular, on the question of perceived trait desirability and its relation to respondent characteristics (e.g. Dohrenwend, 1966; Phillips & Clancy, 1972;

Klassen *et al.* 1975; Gove & Geerken, 1976). These authors agree on the potential for distortion of responses but differ as to the resolution of the problem. Work on health interview surveys for the National Centre for Health Statistics has emphasized 'behaviour interaction' as important in determining the amount of information obtained in an interview, while showing that the psychological and demographic characteristics of respondents have little effect. Choi & Comstock (1975) showed that in the analysis of 15 psychosocial tests the responses obtained by one interviewer differed significantly from those of the remaining 5 interviewers. In the study described here the focus is on the interviewer and the systematic effect of the interviewers on the responses from those they interview.

3.2.1 Description of the study

In order to obtain estimates of interviewer variance for a survey it is necessary to modify the execution of the field operation so that respondents are allocated at random to interviewers. Such a modification of the field work strategy inevitably increases interviewer travel and thus survey researchers are reluctant to undertake investigations of interviewer effect. This West London survey provided a favourable opportunity, however, as a large proportion of the interviews was clustered tightly in urban areas and the randomization of interviews was not expected to increase costs unduly. Thus it was possible to randomize the allocation of 317 addresses, all of which were located in the high noise stratum (Noise and Number Index contour $\geq$ 45), across *eight* interviewers in the study. A complete description of the sample design is given by Wiggins (1979).

The 317 addresses issued produced 307 eligible individuals, of whom 236 (or 77%) responded. The response rates by interviewer are presented in Table 3.5. Three of the interviewers were male and 5 were female; all had had at least 6 months but not more than 2½ years experience of interviewing. Their ages ranged from 34 to 63 years. It was not possible to obtain any additional sociodemographic or psychological data on the group.

Table 3.5: Response rates by interviewer in the noise survey

Interviewer	1	2	3	4	5	6	7	8	Total
Eligible individuals	38	38	40	38	39	39	36	39	307
Interviews completed	26	28	36	29	28	31	28	30	236
Response rate (%)	68	74	90	76	72	79	78	77	77

Interviewer effect was examined by means of 3 major sets of variables. The question numbers below refer to their position on the main questionnaire:[1]

(1) *Noise annoyance as measured* by:

(*a*) an emotional evaluation of the nuisance produced by aircraft noise, a score described by McKennel (1973).

(*b*) an evaluation of how 'bothered' the respondent feels about aircraft noise, which is also part of the above scale score.

(2) *Sensitivity to noise* as measured by:

(*a*) an overall view of the respondent's reactivity to noise (Q. 22).

(*b*) the number of noises mentioned as provoking nuisance (Q. 24).

(3) *Psychiatric morbidity* as measured by the General Health Questionnaire (GHQ) (Goldberg, 1972), a screening instrument administered at the end of the interview.

Although the GHQ is a self-administered questionnaire, it was considered desirable to analyse the responses for the presence of interviewer effect since it was completed in the presence of the interviewer at the end of the interview, the tone of which might well have affected the informants' responses.

3.2.2 Empirical results

In all 41 individual items were analysed in order to estimate the magnitude of the interviewer effect. Table 3.6 presents the distribution of the values of ρ_i (the estimated interviewer effect) over all the items, together with the value of the multiplier $[1+\rho_i\,(m\text{-}1)]$ for each level of ρ_i.

1

Copies of this questionnaire were obtainable from Dr. A. Tarnopolsky, Institute of Psychiatry, De Crespigny Park, Denmark Hill, London SE5 8AF

Table 3.6 Distributions of values of ρ_i for 41 questionnaire items in the noise annoyance survey

ρ_i	No. items	Cumulated no. items	Value of $[1+\rho_i(m-1)]$ (m = 30)
0.00	14	14	1.00
0.01	3	17	1.29
0.02	9	26	1.58
0.03	4	30	1.87
0.04	1	31	2.16
0.05	2	33	2.45
0.06	3	36	2.74
0.07	3	39	3.03
0.08	1	40	3.32
0.09	1	41	3.61

It should be stressed here that the values of ρ_i in the table are estimates of the underlying parameter values and are subject to sampling variability. The distribution of values, however, provides strong evidence of interviewer effect. More than a quarter of the items show values of ρ_i significant at the 0.05 level and 8 items have values of ρ_i significant at the 0.01 level. The last column of the table gives an indication of the potential impact of interviewer effect on these items. The effect of a value of $\rho_i = 0.035$ is to *double* the variance of that item. Thus, if we were to estimate the variance of such an item in the usual way, ignoring the possible error underestimates the true variance by 50%, with serious consequences for the significance of that item.

Figure 3.1 presents the data in graphical form, distinguishing between the items in the *annoyance scale* and the GHQ. From the figure it can be seen that the individual items in the annoyance scale are, in general, more sensitive to interviewer effect than those in the GHQ. We must, however, consider the analytical context in which these items are used. In both cases the primary function of the items is to form components of an additive scale, the total score for which is the variable (or measure) of interest to the researcher. This issue is discussed further in Chapter 5.

Figure 3.1 about here

3.2.3 Summary and conclusion for the noise annoyance survey

This study provides estimates of the impact of variability between interviewers on a set of measures obtained from an epidemiological interview survey. The results given in Table 3.6 and figure 3.1 indicate that for many of the attitudinal items the variance of estimates derived from the survey will be inflated by a factor of between 1.6 and 3.6 due to the presence of interviewer effect. Moreover, for the two scales involved - the GHQ total score and the annoyance score - the effect persists. This analysis suggests two principal conclusions. First, at the developmental (pilot) stage of a survey it would be appropriate to incorporate a similar design of the interviewer allocation so that items sensitive to interviewer effect can be rejected (or modified) before the main field work is carried out. Secondly, more work needs to be done on the relationship between the ρ-values for individual items and the ρ-value obtained on the overall scale score. An analysis of the 10-item annoyance scale is presented in Chapter 5.

3.3 The British Household Panel Survey (BHPS)

The response variance study embedded in this survey compares the relative impact of interviewer effects and sample design effects on survey precision by making use of an interpenetrated PSU/interviewer experiment which was designed by the Colm O'Muircheartaigh and Pamela Campanelli for implementation in the second wave of the British Household Panel Study (BHPS). Section 3.3.1 describes the data and methods used; section 3.3.2 explores the results over all BHPS variables; finally, section 3.3.3 summarises and discusses the findings and their implications for survey research practice.

3.3.1 The BHPS, the interpenetrated design, and the analytic method

The data source for this analysis is the British Household Panel Study (BHPS) which is conducted by the ESRC Centre for Micro-social Change at the University of Essex, UK.

Interviewing on the BHPS began in 1991 and is scheduled to continue in annual waves until at least 2001. The survey used a multistage stratified cluster design covering all of Great Britain. The survey instrument comprised a short household level questionnaire followed by a face-to-face 45 minute interview and short self-completion schedule with every adult in the household. Topics covered include household organisation, income and wealth, labour market experience, housing costs and conditions, health issues, consumption behaviour, education and training, socio-economic values, and marriage and fertility.

An interpenetrated design was implemented in a sample of primary sampling units (PSUs) in Wave II of the survey. Due to field requirements and travel costs, a constrained form of randomisation was adopted in which addresses were allocated to interviewers at random within geographic `pools'; these pools are sets of 2 or 3 PSUs. Every PSU whose centroid was no more than 10 kilometres away from the centroid of at least one other PSU was eligible for inclusion in the design. One hundred and fifty three of the 250 PSUs in the BHPS sample were eligible. Mutually exclusive and exhaustive combinations of these 153 eligible PSUs were formed; this process resulted in 70 pools of PSUs, most with two, and some with three, PSUs each. A systematic sample of 35 pools was then selected for inclusion in the interpenetrating sample design. Great Britain was partitioned for the sample design into 18 regions; only two of these did not include at least one geographic pool.

Of the 35 geographic pools formed four proved to be ineligible as the same interviewer was needed to cover all of the PSUs in the pool and one proved to be effectively ineligible for analysis as one interviewer was needed to cover 3/4 of the geographic pool. An examination of the 30 areas in which the design was implemented does not indicate any systematic abnormality. To the extent that an abnormality did exist, it would affect our results only if it were to interact with the effect of interviewers or with the design effect.

Twenty-five of the thirty usable geographic pools included two interviewers and two PSUs and five included three interviewers and three PSUs. Within PSUs in a given pool, households were randomly assigned to the interviewers working in those PSUs. The sample size for analysis of the 30 geographic pools was 1,282 households and 2,433 individual

respondents.

The initial focus was on the calculation of intraclass correlation coefficients (ρ) for each of the components from the interpenetrated design. These included the interviewer (ρ_i) and the PSU (ρ_s). These coefficients were estimated for all variables in the dataset for which there were 700 or more responses.[2] Categorical and most ordinal variables were transformed into binary variables prior to the analyses; ordinal attitude scales (Likert scales) were, however, treated as continuous. Hierarchical analyses of variance were then carried out for each of these variables using the SPSS MANOVA option. These hierarchical analyses of variance were restricted to cases from the 2x2 geographic pools as the program would not handle the simultaneous calculation of 2x2 and 3x3 geographic pools. The elimination of the 3x3 geographic pools resulted in a reduction in sample size of about 20% at the household level (to 1,010 households) and at the individual level (to 1,903 individuals).

The sums of squares were partitioned using a 'regression approach' in which each term is corrected for every other term in the model.[3] This makes sense substantively and also facilitates the comparison with multilevel modelling using MLn which is described in Chapter 10. It also means that the values for ρ_i and ρ_s which are reported are conditional on each other. Data from the hierarchical analysis of variance runs were then assembled to create a meta dataset of ρ estimates. Other information was added to this dataset such as question type (attitudes, facts, quasi-facts, and interviewer checks) and topic area of the questionnaire.

3.3.2 Findings from Hierarchical Analysis of Variance

The *design effect* is the most commonly used measure of the impact of within PSU homogeneity on survey results; this is deff = $1 + \rho_s(b-1)$ where s denotes the sample clustering, ρ_s is the intra-cluster correlation, and b is the average number of elements selected

[2] In general, the MANOVA analyses which were used required 74 degrees of freedom. A rough rule of thumb to ensure sufficiently stable estimates is to set *n* greater than or equal to the degrees of freedom times 10. Applying this rule to the current models suggests an *n* of approximately 740.

[3] Note that as our design is not balanced, the sums of squares for the various components of the model will not add up to the total sum of squares. Note also that hierarchical analysis of variance assumes a continuous dependent variable. For proportions between .20 and .80, however, the approximation should be fairly close.

from a cluster (the cluster take)[4]. We present the results of this analysis in terms of the intraclass correlation coefficients for interviewers and PSUs. Both measure the within-unit (interviewer or PSU) homogeneity of the observations. Within-PSU homogeneity is a characteristic of the true values of the elements in the population. Within interviewer workloads the homogeneity results from the interaction between the interviewer and his/her respondents; the effect on the variance of an estimate may however be expressed in a form identical to that for the design effect. The *interviewer effect* is inteff = $1+ \rho_i(m-1)$ where i denotes the interviewer, ρ_i is the intra-interviewer correlation and m is the average interviewer workload. The cluster take and the interviewer workload arise as a result of decisions by the designer of the survey; ρ_s and ρ_i are quantities intrinsic to the population structure and to the quality of interviewers respectively. As such the latter are more portable than the variance components themselves; the variance components themselves can of course be calculated once the ρ values are known.

During the past thirty years or so evidence has accumulated about the order of magnitude of both the intra-cluster correlation coefficient and the intra-interviewer correlation coefficient in sample surveys in the US and elsewhere. Though it is impossible to generalize with any confidence,[5] the evidence suggests that values of ρ_i greater than 0.1 are uncommon. Also, as indicated by the means in Table 3.7, the majority of values tend to be less than 0.02 (all of these values are estimates, which accounts for the presence of negative values in Table 3.7). There is also some evidence, although this is mixed, that different types of variables are

---

[4] We are aware that the approximations involved in ignoring some of the complexities of the sample design affect the estimates produced by MANOVA. In a national sample the available stratifiers (primarily region and measures of urbanization) produce quite modest gains in precision, typically less than 5%. The complexity of adding these to the analyses did not justify their inclusion. Ignoring them may inflate the estimates of ρ_s but only if the effect is greater on the estimate of between cluster variance than on the estimate of the total variance. In the survey literature, we refer to an estimate of ρ that may include the impact of stratification as a "synthetic" ρ.

Selection with pps followed by selection with probabilities inversely proportional to size within clusters will not affect the results though it does affect slightly the interpretation of ρ: ρ is now the estimated intra-cluster correlation coefficient for the subclusters created by the subsampling procedure. As the subsampling procedure for BHPS was essentially random within cluster, this distinction is unimportant here.

[5] There is some difficulty in comparing across studies as each involves different numbers of interviewers, different sample sizes, and different types of variables. In addition, some authors report the negative values of ρ_i which occur and other set these to 0.

affected by interviewers in different ways; attitude items and complex factual items are considered more sensitive to interviewer effect than simple factual items (see, for example, Collins and Butcher, 1982; Feather, 1973; Fellegi, 1964; Gray, 1956; Hansen et al, 1961, and O'Muircheartaigh, 1976 and section 3.2.2).

Table 3.7: Summary of Some Other Interviewer Variance Investigations

Study	Values of ρ_i	Mean
Neighbours noise/illness (UK)		
Gray (1956)	-0.018 to 0.10*	0.015*
TV habits (UK)		
Gales and Kendall (1957)	(0.00) to 0.05, 0.19**	NA***
Census (US)		
Hanson and Marks (1958)	-0.00 to 0.061**	0.011**
Blue Collar Workers (US)		
Kish (1962)		
First study	-0.031 to 0.092	0.020
Second study: Interview	-0.005 to 0.044	0.014
Second study: Self-Completion	-0.024 to 0.040	0.009
Census (Canada)		
Fellegi (1964)	(0.00) to 0.026	0.008
Health Survey (Canada)		
Feather (1973)	-0.007 to 0.033	0.006
Mental Retardation (US)		
Freeman & Butler (1976)	-0.296 to 0.216	0.036
Aircraft Noise (UK)		
O'Muircheartaigh & Wiggins (1981)	(0.00) to 0.09	0.020
Consumer Attitude Survey (UK)		
Collins & Butcher (1982)	-0.039 to 0.119	0.013
9 Telephone Surveys (US)		
Groves and Magilavy (1986)	-0.042 to 0.171	0.009

* Calculated from F-Ratios using formula supplied by Kish (1962)
** Numbers available through Kish (1962)
*** Mean can not be computed. Paper does not report all variables analyzed

The range of values reported in the literature for ρ_s is similar to that for ρ_i, though we would expect ρ_i to have more values near zero. Again, the evidence suggests that values greater than 0.1 are uncommon and that positive values are almost universal. The large values tend to be for certain types of demographic variables, notably tenure and ethnic origin. This is to be expected since adjacent groups of houses in a small area will tend to be of similar type and tenure (Lynn and Lievesley, 1991). Other demographic variables such as sex and marital status tend to show very low values. It is typically found that behavioral and attitudinal variables have ρ_s values somewhere between these extremes, with attitudinal variables showing slightly lower values than behavioral ones. In the World Fertility Survey (see Verma, Scott, and O'Muircheartaigh, 1980), the median ρ_s across various countries was 0.02 for various nuptiality and fertility variables. The median was much higher (around 0.08) for variables concerning contraceptive knowledge.

In comparing these two sources of variability, Hansen, Hurwitz and Bershad (1961) found that interviewer variance was often larger than the sampling variance. Bailey, Moore, and Bailar (1978), on the other hand, found response variance components that were at least 50% of their sampling variance for only a quarter of their statistics.

Figure 3.2 about here

We included in the analysis 820 variables from the BHPS. Of these, 98 were attitude questions, 574 were factual, 88 were interviewer checks (items completed by the interviewers without a formal question), and 60 were quasi-facts (mostly on a self-completion form). Figure 3.2 shows the cumulative frequency distributions for ρ_s and ρ_i. The orders of magnitude for the two coefficients were strikingly similar. As these values are themselves estimates they are subject to imprecision; using a test of significance at the 5% level 4 in 10 of the values of ρ_s and 3 in 10 of the values of ρ_i were significantly greater than zero. In the case of ρ_s this is not surprising as positive values are expected for most survey variables. What is somewhat surprising is that, within the study, ρ_i is of the same order of magnitude. For these data, because of the way the investigation was designed, the average interviewer workload and the average cluster take were the same; thus our estimates of ρ_s and ρ_i imply

that the *impact* of the sample design and the interviewers were also about the same.

All types of questions show the presence of significant values of ρ_i, using an F test comparing between group variances with the error variance. For attitude questions, 28% of the values of ρ_i were significantly greater than zero; for factual questions it was 26%; for interviewer checks, a staggering 58%; and for the quasi-factual questions, 25% (with the exclusion of the self-completion items). What is interesting is the similarity of the findings for the attitudinal and factual items, which is in contrast to the findings of some studies. There is some variation between types of attitudinal item. Among those items based on Likert scales, 33% showed significant values of ρ_i; this compares with 25% of the other attitude items.

We also looked for differences by source of the question. For example, 32% of the items in the individual schedule had ρ_i values which were significantly greater than zero. The same was true for 17% of the self-completion items, 27% of the coversheet items, 28% of the derived variables from the individual's questionnaire, 32% of the household questionnaire items, and 34% of the derived variables from the household questionnaire. The notable difference here in susceptibility to interviewer effects is between the self-completion items and those that are interviewer-administered. The fact that there is an interviewer effect at all on the self-completion form is interesting. Kish (1962), for example, found little evidence to suggest the presence of such an effect on the written questionnaires he examined. O'Muircheartaigh and Wiggins (1981 and section 3.2), however, did find an effect for a health supplement completed in the presence of the interviewer (as were the BHPS self-completion items).

There was also basically no difference in the proportion of significant ρ_i values between the different sections of the questionnaire: demographics, health, marriage and fertility, employment, employment history, values, and income and household allocation (with the percentage significant ranging from 22% to 35%). In contrast the section at the end of the questionnaire for interviewers to record their observations was highly susceptible to interviewer effects. Seventy-six percent of the items in the interviewer observation section showed significant values of ρ_i. There was also a difference between dummy and continuous

variables, with a higher proportion of effects being noted for the continuous variables.

Furthermore, there was a clear positive correlation of 0.35 between ρ_i and ρ_s. A positive correlation between ρ_s and ρ_i implies that variables that show large intra-cluster homogeneity (show relatively substantial clustering among true values) are also sensitive to differential effects from interviewers. Such a correlation has not, to our knowledge, been observed before. As the elements in the computation of this correlation are themselves variables, the absence of such evidence may be because it is necessary to have a large number of variables to estimate such a correlation coefficient with any precision. In our analysis the correlation shows striking consistency across types of variables.

Homogeneous clusters contain individuals similar to one another; it is not unreasonable to suggest that individuals with similar values on the variable in question may respond in a similar way to whatever qualities the interviewer brings to bear in the interviewer-respondent interaction. This would mean that variables that manifested intracluster homogeneity would on balance be more likely than other variables to be display intra-interviewer homogeneity. An alternative explanation may be found in some of the early work on interviewers (see Hyman, 1954). Interviewer expectations are known to influence the responses obtained by interviewers. For a variable to have a relatively large value of $\bar{\rho}_s$ the individuals within a cluster will have relatively homogeneous values; it is possible that this consistency will affect the interviewers' expectations as the interviewer's workload progresses, leading to enhanced correlations within interviewer workloads.

This latter explanation is consistent with the technical interpretation of the correlation between the response deviation and the sampling deviation for a single variable postulated in the Census Bureau model and included in Hansen et al (1961), Fellegi (1964), and Bailey et al (1978). It is not possible to estimate this correlation directly for a single variable without at least two waves of data collection, though it is included in the standard model estimate of ρ_i. Hansen et al (1961) give an example of how this latter correlation may arise for a single variable.

3.3.3 Summary and discussion for the BHPS study

The assumption underlying most statistical software - that the observations are independent and identically distributed (*iid*) - is certainly not appropriate for most sample survey data. Variances computed on this assumption do not take into account the effects of survey design (eg inflation due to clustering) and execution (eg inflation due to correlated interviewer effects).

There are two different reasons why we might be interested in interviewer effects and sample design effects. The first is to establish whether the sample design (typically clustering in the design) and/or the interviewer (because many respondents are interviewed by each interviewer) have an effect on the variance-covariance structure of the observations. This is the traditional sample survey approach and includes consideration of the *design effect* and the *interviewer effect* following the ANOVA and Census Bureau models. The emphasis is on the estimation of means or proportions and on the standard errors of these estimates; variance components models do not add anything to these analyses.

This analysis with a specially designed study in wave II of the British Household Panel Survey (BHPS) permitted us to assess both these inflation components. Across the 820 variables in the study, there was evidence of a significant impact of both the population clustering and the clustering of individuals in interviewer workloads. The intraclass correlation coefficient, ρ, was used as the measure of homogeneity. We found that sample design effects and interviewer effects were comparable in impact, with overall inflation of the variance as great as five times the unadjusted estimate. The median effect across the 820 variables was an 80% increase in the variance. The magnitude of the intra-interviewer correlation coefficients was comparable across these types, though the most sensitive items tended to be the interviewer check items. There was a tendency for variables that were subject to large design effects to be sensitive also to large interviewer effects and we offer a possible interpretation of this correlation in section 3.3.2.

The large values of ρ_i on particular items and the fact that ρ_i is of the same order of magnitude

as ρ_s suggests that survey organisations should attempt to incorporate measurement of ρ_i into their designs. If the necessary modifications of the survey design are too expensive to allow this, organisations should at least try to minimise its impact; this could be accomplished by reducing interviewers' workloads. Current practice tends to favour smaller dedicated interviewer forces with large assignments; in the presence of substantial interviewer effects this is a misguided policy.

The second reason we are interested in these correlated variances is to ensure that effects on the univariate distributions do not contaminate our estimates of relationships among variables in the population; in this case our objective is to control the effects or to eliminate them from the analysis. The standard approach of the survey sampler is to estimate the parameters assuming *iid* and produce design-based variance estimates using re-sampling methods such as the jackknife or bootstrap; this however is only an approximate solution. The explicit modelling of effects is both more precise and more informative. In this situation there are two aspects of interest: whether explicitly including the sample clustering and the interviewer workloads in the model changes the estimates of the relationships (the contamination issue), and whether the clustering and interviewers have an effect on the distribution of values obtained for the dependent variable. This topic is pursued in Chapter 10.

3.4 Conclusion

The three disparate studies reported in this chapter provide a general picture of the magnitude and impact of the interviewer effect (defined as the correlated variance due to interviewers) on the survey results. Each of the data sets is analyzed further in later chapters. From the absenteeism study the effect of the interviewer on summative attitude scales is explored and a multivariate analysis of interviewer effects is proposed (Chapter 5). The noise annoyance survey provides more material for the analysis of scales, and offers an opportunity to incorporate the interviewer directly into substantive analysis (Chapters 5 and 10). The BHPS data form the basis for a multilevel analysis of interviewer effects (Chapter 10).

Chapter 4 INTERVIEWER VARIANCE IN THE WORLD FERTILITY SURVEY

This chapter presents the design and general results for a major project designed and carried out on behalf of the World Fertility Survey (WFS) between 1978 and 1984. In each participating country the study consists of a single-round survey based on a probability sample of households. Though the sample in each case is designed individually to suit the country's situation, all the samples were designed to be *measurable*, ie the design permits the estimation of sampling errors from the survey data themselves. As a matter of policy, estimates of sampling errors were computed as part of the first stage of analysis. A full discussion of WFS sample characteristics may be found in Verma, Scott and O'Muircheartaigh (1980).

In the context of the WFS, methodological experimentation is by and large excluded by the very nature of the operation. The primary objective has been to assist countries in obtaining the best possible data from a single operation, which necessarily requires the choice of a study design considered *a priori* to be the most suitable. Thus it has not been possible in general to compare different survey procedures in order to ascertain which is superior. Furthermore, for the data collected in WFS surveys, there is no source of external validation data available at the level of the individual respondent. Consequently the analysis of response errors must be based on an examination of the internal consistency of the data. The analysis therefore is of the *reliability* of the data, rather than of their validity.

4.1 The Structure and Design of the Project

The WFS programme provided a unique opportunity for a cross-national study of response error. The diversity of the WFS participant countries permits the selection of study areas covering a considerable regional, developmental, ethnic and cultural spread, while the standardization of the survey instruments and the centralization of technical control make it possible to achieve meaningful comparison between the national studies. Moreover, the main surveys were already financed and only the marginal costs of the operations and modifications directly attributable to the response errors project needed to be found.

The main emphasis of the project was twofold. First, the observed variability of the results was to be partitioned into the components representing *sampling variance* (sampling error) and the *simple response variance*; and secondly, the magnitude of the *correlated response variance* due to the interviewers was to be estimated and its impact assessed (this component is frequently known as *interviewer variance* or *interviewer effect*).

The project also provided an opportunity to gain some insight into the way in which the individual responses vary from one occasion to another for the individual. The cross-tabulation of the responses for the main survey interview and the re-interview was expected to provide a detailed picture of the *stability* of the responses.

It was decided that the project should be carried out in four countries representing a wide range of developmental levels, each selected from a different broad cultural area. The four countries originally chosen were Lesotho, in Africa; Peru, in South America; Turkey, in the Middle East; and the Philippines, in Asia. Due to difficulties in implementation, the project was actually carried out in full only in Lesotho and Peru.

The main features of the project design were the same for each of the countries, although the detailed implementation differed in each case. The overall design consists of two stages. The first involves a modification of the design of the main survey. The other was separate from and additional to the main survey.

Stage 1 The main survey as planned for the country forms the basis of the first stage. The only modification is that a subsample of the respondents is allocated randomly to each of the interviewers in the team in each location. The randomization is carried out at the survey headquarters before the fieldwork commences in that location. The cost of this stage arises from the additional clerical work involved in designating the subsample to be interviewed by each interviewer; and from the possible increase in the cost of the fieldwork due to the inflexibility of interviewer allocation, since the freedom of decision of the supervisor is restricted by specifying in advance which interviewer should carry out each interview.

Stage 2 The second stage consists of a re-interview with at least a subsample of the respondents from the main survey. The re-interviews should be carried out under the same essential survey conditions as the original interviews. Furthermore the same team should do the original interviews and the re-interviews in each location selected for the second stage. No interviewer should interview the same respondent twice; the re-interview should be carried out by another member of the same team. The questionnaire used in the re-interview should be the same as that used in the original interview in so far as this is possible. Cost considerations demand that the re-interview survey should be carried out only for a subsample of the main survey. It is desirable that this subsample should be a probability sample of the main survey sample.

For the re-interview survey, the basic problem is that of timing. If the time period between the main survey and the re-interview is short, the responses will not be independent and may indeed be highly correlated. The first interview may affect the second response if, for example, the respondent remembers the answers given in the first interview. This problem may be avoided, or at least reduced, by leaving a long time period between the two interviews, but this would lead to two further problems, namely that genuine changes may occur in the interval, which will exaggerate apparent discrepancies between the two responses, and that the interviewers used in the main survey may no longer be available. On balance the advantage lies with choosing a short interval between the two stages, particularly because of the need to have the same interviewers in both surveys.

Sections 4.1.1 and 4.1.2 below describe the procedures used in Lesotho and Peru .

4.1.1 Lesotho

The Lesotho Fertility Survey, conducted by the Central Bureau of Statistics in 1977-8, was based on a national two-stage probability sample. Census enumeration areas (of which there are 1066 in the country) were the primary sampling units (PSUs). One hundred PSUs were selected with probabilities proportional to size and a sample of households was selected within

each selected PSU such that each household in the population had an equal probability of selection. The PSUs were stratified by ecological zone, population density, and size before the first stage of selection. All ever-married women residing (on a *de facto* basis) in the selected households were eligible for interview. In all, 3603 individual interviews were successfully completed, giving an overall response rate of around 88 per cent.

Fieldwork for the main survey was carried out by eight teams of interviewers, each consisting of either four or five interviewers, one supervisor and one field editor. In all, 34 interviewers worked on the survey. The language in which the interviews were conducted was Sesotho. The questionnaire itself was also in Sesotho, although the interviewers' instruction manuals were in English.

Arrangements were made for the interpenetration (randomization) of the interviewer workloads within the teams for every PSU in the sample. For each PSU, the selected sample of households was listed, village by village, in the order in which the fieldwork was to be carried out. The numbers 1 to 5 (for teams with five interviewers) or 1 to 4 (for teams of four interviewers) were allocated to each successive set of five or four households on the list. For each of the numbers, a separate list of the households with that number was written out. For each team one of the lists was allocated at random to each interviewer in the team, before the fieldwork began. The supervisor received the master list and the set of interviewer lists for each cluster (PSU) in the team's work allocation, together with a list giving the allocation of workloads to interviewers. The supervisor was given the responsibility of ensuring that each interviewer carried out all her own workload.

In deciding on the subsample to be selected for the re-interview survey, two alternative strategies were considered. The first strategy was to use all eight teams in the re-interview survey and to have each team complete a part of its previous work allocation. The difficulty with this approach, however, lay in the fact that each team required a vehicle to carry out its fieldwork and vehicles were available for all teams only for the period of the main fieldwork (August - October 1977). Only three vehicles could be obtained for the period October - December 1977. Thus the second strategy was adopted; three teams were chosen for the re-

interview survey and each of these was allocated two-thirds of the PSUs in which it had worked in the main survey. Each team was assigned an additional female field editor for the re-interview survey. These field editors were chosen from those who had worked in other teams in the main survey.

The system of allocation of workloads to interviewers in the re-interview survey is given below. The allocation is given for teams of four and five interviewers.

No of interviewers in the team	4	5
Interviewers for main survey	1 2 3 4	1 2 3 4 5
Interviewers for re-interview	4 3 2 1	5 4 3 2 1

Where interviews were, for any reason, re-assigned for the original interview, the allocation for the re-interview was re-assigned accordingly.

The questionnaire for the main survey in Lesotho incorporated the WFS core questionnaire and two modules. The questionnaire for the re-interview was shorter, consisting of sections 1, 2, 3 and 5 of the core questionnaire. All the questions asked in the re-interview had already been asked, in exactly the same form, in the original interviews. The questionnaires from the original interviews were edited and coded in advance of the re-interview survey. The re-interviews were edited and coded in the field by the supervisory staff and were compared by the editors for a subset of the questions. Where discrepancies were found between the answers to these questions, a third, reconciliation, interview was to be carried out by one of the female supervisory staff. These reconciliation interviews were to be carried out before the team left the area. The editors were instructed to use a clean questionnaire, fill in the identification information, and mark the questions to be reconciled (ie the questions where inconsistencies were found). A special summary form was prepared for each reconciliation interview.

Implementation

In Lesotho the execution of the project design in the field conformed closely to that described above. The fieldwork for the main survey lasted from August to early October 1977. The re-interview survey commenced in late October and was completed in December 1977. One of the interviewers left the field staff between the two field operations and was replaced by an interviewer from one of the teams not involved in the re-interviews. The time interval between the two interviews varied between one month and four months. Twenty-five PSUs were included in the re-interview survey and a total of 724 interviews were obtained from the 867 individuals, a response rate of 84 per cent. The system of allocation of workloads to interviewers in the two field operations was implemented satisfactorily. One additional benefit obtained from the response errors project may be noted here. On examining the field records for the main survey, there appeared to be too many cases where the code `dwelling vacant' had been obtained as the final response category. Since it seemed possible that this code had been misunderstood by the interviewers, it was decided to check the dwellings with this code in a number of PSUs during the re-interview survey fieldwork. Of a total of 62 such cases in the 15 PSUs which were checked, 26 (or 42 per cent) produced completed interviews. These cases provided both an opportunity to improve the data from the main survey and an indication of the possible impact of such non-response on the results of the main survey.

4.1.2 Peru

The Peru Fertility Survey, conducted by the National Statistics Office during 1977-8, was based on a three area-stage national probability sample. Districts (of which there are around 1700 in the country) formed the primary sampling units (PSUs). In all, 124 PSUs were selected, 57 self-representing, appearing in the sample with certainty, and 67 non-self-representing, selected with probability proportional to size. In urban areas *blocks*, and in rural areas *localities* constituted the second-stage units (SSUs). Generally, an SSU consisted of 25-100 dwelling units, and a total of 1424 SSUs were selected with probability proportional to size. The third sampling stage involved the systematic selection of dwellings from within the selected SSUs, yielding a self-weighting sample (except that jungle areas were oversampled by a factor of 4). All ever-married women aged 15-49 residing (on a *de facto* basis) in the 8330 sample dwellings were eligible to be interviewed in detail regarding their

maternity and marriage histories, knowledge and use of contraception, fertility intentions and preferences and socio-economic background. In all 5640 individual female interviews were successfully completed, representing a response rate of around 90 per cent.

Fieldwork for the main survey was to be conducted by 36 female interviewers divided into six teams, each team working under one supervisor and one field editor. It was necessary to use five different languages or dialects for interviewing: Spanish; Aymara; and three Quechua dialects, Ancash, Ayacucho and Cuzco. Arrangements were made in the main survey for the interpenetration (randomization) of the interviewer workloads within teams for the secondary sampling units (SSUs) selected for the response errors project, the designated SSUs. For each SSU a folder was prepared containing the basic information about the SSU and listing the selected households. Each team was given a set of these folders before going into the field. For each SSU a decision was taken as to how many interviewers should be sent to the SSU. At least two interviewers were to travel to each SSU and the interviews were to be allocated randomly between them; the maximum number of interviewers in a team was seven.

In urban PSUs, particularly Lima, the allocation of interviewers to households was carried out over the whole designated sample. If, for example, a PSU contained five SSUs and the team contained seven interviewers, the letters A to G were allocated to each successive set of seven households in the PSU as follows:

SSU 1 9 households A B C D E F G A B
SSU 2 10 households C D E F G A B C D E
SSU 3 7 households F G A B C D E
SSU 4 15 households F G A B C D E F G A B C D E F
SSU 5 4 households G A B C

Each interviewer was allocated randomly one of the letters A to G and the households bearing that letter constituted the interviewer's workload.

In rural areas and where the number of households in a group of designated SSUs was too small, an appropriate subset of the letters A to G was to be used, eg if there were only four

households, the letters A, B, C, D, or A, B, A, B would be allocated. Each letter would identify one of the interviewers sent to the SSU.

Approximately one in four of the main survey SSUs (urban blocks and rural localities) were designated for the Response Errors Study (RES). The RES consisted of conducting a *re-interview* with all respondents in the designated SSUs, using a shortened but otherwise identical version of the original questionnaire. Following this, the completed questionnaires for the first and the second interviews were compared by the field editors, and in cases where major inconsistencies occurred a third, *reconciliation*, interview was carried out to ascertain the `true' response and also the cause of the discrepancy.

In Lima the designated SSUs were selected at random and the re-interview involved a separate trip to the selected areas. Outside Lima, owing to more difficult travel, the sample was selected purposively, and fieldwork logistics were planned such that while covering a group of neighbouring SSUs for the original interview the team would pass through the designated cluster(s) twice, with an interval of 1-2 months between the two visits. Figure 4.1 below illustrates the principle. Starting from SSU 1 (say, a district centre), a team conducts the first interview in 10 clusters and conducts re-interviews in the purposively selected clusters 2 and 4 during its return trip.

Figure 4.1 here

A rotating system of allocating workloads to interviewers for the first and second interviews was devised and is given below. The allocation is presented for the maximum team size of seven interviewers and for each subset of interviewers. If, for any reason, any interviewer should fail to carry out any part of her workload, or if an interviewer should complete an interview allocated to another, this fact, the reasons for it, and the names and numbers of both interviewers was to be recorded by the supervisor and reported to headquarters at the end of the fieldwork.

The questionnaire for the main survey in Peru incorporated the WFS core questionnaire and

the fertility regulation module. The questionnaire for the re-interview was shorter but all the questions included had already been asked during the original interview.

In Lima the questionnaires from the main survey were edited and coded in the survey headquarters before the re-interviews were carried out. The completed questionnaires were *not* shown to any of the interviewers before the re-interviews. The completed questionnaires for the two interviews were compared by the editors for all the questions asked in the re-interview. When inconsistencies were found, a reconciliation interview was carried out to ascertain, if possible, the cause of the discrepancy. In rural areas, the completed questionnaires were kept in the custody of the supervisor/editor and the reconciliation interview was carried out before the team left the SSU.

Rotating System of Allocation of Workloads

No of interviewers in the area	2	3	4	5	6	7
Interviewers for main survey	A B	A B C	A B C D	A B C D E	A B C D E F	A B C D E F G
Interviewers for re-interview	B A	C A B	D C B A	E C D B A	F E D C B A	G F D E C B A

Implementation

The study design required that (i) at least two interviewers travel together to an SSU with interviews within the cluster allocated randomly between the interviewers, and that (ii) re-interviewing in the cluster should be done by the same team, following a predetermined random allocation such that no respondent is interviewed twice by the same interviewer. For several reasons the pattern of interview allocation diverged rather substantially from that planned. The primary reason was the disruption of the implementation of the main survey, due to climatic and budgeting problems, resulting in the fieldwork being stretched over a very long period. Consequently the time elapsed between the two interviews also tended to be lengthened: while 60 per cent of the re-interviews were conducted within three months of the

original interview, the time elapsed exceeded six months for nearly 30 per cent. This made it difficult in practice to follow the above-mentioned allocation rules. Further, urban and rural areas differed greatly (not unexpectedly) in relation to the elapsed time: nearly 80 per cent of the re-interviews in urban areas but only 30 per cent of those in rural areas were conducted within three months of the original interview; the interval exceeded six months for only 5 per cent in urban areas, but for nearly 70 per cent in rural areas (there being very few re-interviews in rural areas between the fourth and sixth months). This disrupted the plan to conduct re-interviews in rural areas during the return trip. It is noteworthy, nevertheless, that an overall response rate of around 85 per cent was achieved in the re-interview survey. Another difficulty resulted from the rather small sample take per SSU (an average of around four, not infrequently only one or two interviews per cluster). It was not always possible to send two interviewers to each cluster.

Though an attempt was made to achieve a reasonable geographical spread in purposively designating re-interview areas, the resulting re-interview sample none the less differed significantly in composition from the main survey sample. There was an over-representation in the former of urban areas, as well as of better educated women. Since both these characteristics are likely to be strongly related to response errors, it was necessary to weight the re-interview sample so that its *joint* distribution by city size (four categories) and women's level of education (five categories) agreed with the main survey sample. The range of weights introduced was around 1-5.

4.2 The Correlated Response Variance in Lesotho and Peru

In Lesotho and in Peru the overall design included both random allocation of respondents to interviewers and re-interviewing of the respondents. The estimation here is confined to the correlated response variance due to interviewers - the interviewer variance. The interviews from the main survey and those for the re-interview survey are analysed separately; the magnitude of the correlated response variance can be estimated separately for the same set of variables in each case.

The correlated response variance - in this case the *interviewer variance* - is:

$$\frac{\rho_I \sigma_\varepsilon^2 (m - 1)}{n} \tag{4.1}$$

As was the case for the examples in chapter 3, we cannot, for each set of interviews, estimate ρ_I directly. We can, however, estimate $\rho_I \sigma_\varepsilon^2$. A good indicator of the potential impact of the interviewer variance is

$$\rho_I I = \rho_I \sigma_\varepsilon^2 / \left(\sigma_y^2 + \sigma_\varepsilon^2\right) . \tag{4.2}$$

where I is the index of inconsistency. This of course is ρ_i , the intra-interviewer correlation coefficient defined in chapter 2 and used in chapter 3.

Table 4.1L presents the estimated values of ρ_i for the variables for which significant results were found in Lesotho. The estimation procedure provides three separate estimates of ρ_i -one for the whole of the main survey and one for each phase of the Response Errors Project for the respondents who were interviewed twice. The latter two sets of estimates are based on distinct data sets, the first for the interviews from the main data collection, the second from the reinterviews.

Table 4.1L Estimates of ρ_i for Lesotho

Variable	Main survey (total sample)	Main survey (designated sample)	Re-interview survey (designated sample)
Ever-use of contraception	0.084	0.122	0.119
Years of education	0.015	0.077	0.077
No of children desired	0.041	0.041	0.048
First birth interval	0.045	0.163	0.084
Martial duration	0.000	0.016	-0.005
Age at marriage	0.005	-0.006	0.032

The values in table 4.1L are all estimates and are themselves subject to variance. It is noteworthy, however, that the same four variables emerge as the most sensitive to interviewer effect in all three analyses. The estimates in the second column are based on a subset of the responses considered in the first column, but the third column represents an entirely different set of responses. The result for *Years of education* for the total sample is surprisingly low, but otherwise the results provide reassurance on the representativeness of the designated sample.

The last two variables are included in the table as one (marginally) significant value was obtained in each case. The weight of the evidence suggests, however, that neither variable is actually subject to interview effect but that the value is merely the result of chance variation.

Table 4.1P gives the estimated values of ρ_i for the five variables that produced significant results for Peru. Our estimation procedure provides separate estimates of ρ_i for the two phases in Peru, and although again these values are all estimates it is interesting to note that with one exception the same five variables emerged in both analyses as those most sensitive to interviewer effect. The exception is *Ever-use of contraception* which had the largest estimated interviewer effect in the main survey, an effect which disappears completely in the re-interviews.

Table 4.1P Estimates of ρ_i for Peru

Variable	Main survey	Re-interview survey
Ever-use of contraception	0.10	0.00
Whether worked	0.04	0.05
Education	0.06	0.05
No. of children desired	0.03	0.15
First birth interval	0.02	0.04

The values of ρ_i (or $\rho_1 I$) provide an index of the susceptibility of variables to interview effect. The magnitude of the variance component may be expressed either as

$$\frac{\rho_I \sigma_\varepsilon^2}{n} (m - 1) \tag{4.3}$$

or alternatively using as ρ_i (or $\rho_I I$)

$$\rho_I I \ (m-1) \ \frac{\left(\sigma_y^2 + \sigma_\varepsilon^2\right)}{n} \tag{4.4}$$

which has the advantage that it uses as a base the value of

$$\frac{\left(\sigma_y^2 + \sigma_\varepsilon^2\right)}{n}$$

which is the simple total variance. The simple total variance is easily and directly estimable from the survey data and also provides the base against which the sampling variance is measured in most survey work.

Whichever form is used, the most important point to note is that the average interviewer workload m is critical in determining the magnitude of the variance component. Even a relatively small value of $\rho_I I$ will have a considerable impact on the total variance if the value of m is large. With a value of $\rho_I I = 0.02$, for example, and $m = 100$, the effect of the correlated response variance would be to increase the total variance by an amount equal to twice the simple total variance.

A large of value of ρ_1 would not in itself be sufficient to imply a large increase in the total variance. The size of the simple response variance $\left(\sigma_\varepsilon^2\right)$ is also important. If σ_ε^2 is small

--in particular if it is small relative to the simple total variance -- even a large value of ρ_1 will have little impact. In chapter 6 we investigate the values of σ_ε^2 .

The central point is that the correlated response variance is an additional contribution to the total variance due to intercorrelations between the response deviations. Thus, in principle, if a variable is not subject to fluctuations in response - if there is no simple response variance - there cannot be any correlated response variance. Similarly if the simple response variance is very small, a very high degree of intercorrelation among the response deviations would be

necessary before the correlated response variance could make a substantial contribution to the total variance. If however for a variable with non-negligible simple response variance, the response deviations are sensitive to the behaviour or other characteristics of the particular interviewer who conducts the interview, then the interviewer variance may be an extremely important component of the total variance and could in some cases dominate the total variance.

4.3 Partitioning the Total Variance

Before presenting the results for Lesotho and Peru, it is necessary to set out the form of the partitioning in some detail. The material in this section can be summarized in a simple figure (figure 4.2).

The figure is essentially a 2x2 table, and the cells of the table show the four components of the Total Variance (TV) of the survey estimate as defined in this framework. The two rows represent the sources of variation - the sampling process and the measurement process. The two columns show the two variance types - simple variance arising from (the) uncorrelated (parts of) deviations, and correlated variance arising from correlations among these deviations. The column marginal totals are the simple total variance (STV), the sum of the simple sampling variance (SSV) and the simple response variance (SRV), and the correlated total variance (CTV), the sum of the correlated sampling variance (CSV) and the correlated response variance (CRV). The row marginal totals are the total sampling variance (TSV), the sum of the simple sampling variance and the correlated sampling variance, and the total response variance (TRV), the sum of the simple response variance and the correlated response variance.

Figure 4.2 The Total Variance of the Estimator

Type / Source	Simple	Correlated	Total
Sampling	Simple sampling variance **SSV**	Correlated sampling Variance **CSV**	Total sampling variance **TSV**
Response	Simple response variance **SRV**	Correlated response variance **CRV**	Total response variance **TRV**
Total	Simple total variance **STV**	Correlated total variance **CTV**	Total variance **TV**

The implications of each of these components and of each combination is different. At the heart of the data are the two simple variance components, the simple response variance and the simple sampling variance.

The simple sampling variance is the variance of a simple random sample from the population in the absence of any measurement (response) error, and is essentially a function of the overall variability in the population for the variable being considered, that is, the extent to which individuals in the population differ from one another.

The simple response variance is the variance contributed to the estimate by the whole variety of perturbations and disturbances introduced in to the data by all the factors that can cause the observation to differ from the `true value'. This will include not only an agent such as the interviewer but also transient factors such as timing, the mood of the respondent. The simple response variance is a measure of the intrinsic quality (reliability) of the data.

The correlated sampling variance is the sum total of the effects of the aspects of the sample design that induce the selection of elements that have intercorrelations among them. Stratification, by forcing the inclusion of elements from different strata, will produce a sample with negative inter-element correlations relative to the population a whole; the selection of clusters of elements on the other hand will produce a sample with positive inter-element correlations compared to the elements in the population as a whole. On balance the empirical evidence suggests that these positive inter-correlations overwhelm the negative in practice leading to an overall positive correlated sampling variance. The larger the number of elements selected from a cluster the greater the impact.

Similarly with the correlated response variance the factors that induce positive correlations dominate. In particular the interviewer, each of whom typically interviews (far) more than one respondent, has a consistent impact on the respondents s/he interviews. To the extent that there is flexibility (or "malleability") in the responses the interviewer may tend to mould the response consistently or push the response (or the respondent) in a particular direction, thereby introducing positive within-interviewer between-respondent correlations. The more respondents interviewed by each interviewer the more scope there is for this effect and the greater it will tend to be.

The simple total variance can be estimated directly from the survey data and is:

$$\frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} \tag{4.5}$$

As part of the routine analysis of the WFS surveys, the sampling variance was estimated using the CLUSTERS program. The estimate is actually an estimate of the simple total variance plus the correlated sampling variance. In the notation of this section it is an estimate of:

$$\frac{\sigma_y^2}{n}\{1 + roh\ (b - 1)\} + \frac{\sigma_\varepsilon^2}{n} \tag{4.6}$$

where *roh* is the intracluster correlation coefficient and b is the mean cluster `take'. The correlated sampling variance is

$$\frac{\sigma_y^2}{n} \{roh\ (b - 1)\} .$$

The interviewer variance is

$$\frac{\sigma_\varepsilon^2}{n} \{\rho_1\ (m - 1)\} . \qquad (4.1)$$

The total variance is:

$$\frac{\sigma_y^2}{n} + \frac{\sigma_\varepsilon^2}{n} + \frac{\sigma_y^2}{n} \{roh\ (b - 1)\} + \frac{\sigma_\varepsilon^2}{n} \{\rho_1\ (m - 1)\} . \qquad (4.7)$$

All the components of variance are affected by the sample size, but their relative magnitudes are not dependent on the sample size. Of the factors in (4.7) only two (apart from n) are subject to manipulation through the survey design. These are the interviewer workload size m and the average cluster `take' b.

In order to simplify the presentation some manipulation of the terms used in the earlier sections is required, particularly for the components of the correlated variance. Instead of using σ_y^2/n as a base for the correlated sampling variance and σ_ε^2/n as a base for the correlated response variance it is possible to use the simple total variance $\left(\sigma_y^2 + \sigma_\varepsilon^2\right)/n$ (4.5) as a base for both.

Thus the correlated sampling variance which has previously been written as

$$\frac{\sigma_y^2}{n} \{roh\ (b - 1)\}$$

can alternatively be written as

$$\frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} \cdot \frac{\sigma_y^2}{\sigma_y^2 + \sigma_\varepsilon^2} \{roh\ (b - 1)\}$$
$$= \frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} (1 - I)\{roh\ (b - 1)\} \qquad (4.8)$$
$$= \frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} \rho_s\ (b - 1)$$

where ρ_s is a synthetic intra-cluster correlation coefficient which takes into account the presence of the simple response variance. The quantity *roh* which is estimated by standard sampling error programs is in fact ρ_s and not the pure *roh* in (5.13). The estimate of the design effect, deff, is in fact an estimate of $1+\rho_s(b-1)$.

Similarly the interviewer variance component can be expressed either as

$$\frac{\sigma_\varepsilon^2}{n} \{\rho_I\ (m - 1)\} \qquad (4.1)$$

or, using $\left(\sigma_y^2 + \sigma_\varepsilon^2\right)/n$ as a base, as

$$\frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} \cdot \frac{\sigma_\varepsilon^2}{\sigma_y^2 + \sigma_\varepsilon^2} \{\rho_I\ (m - 1)\}$$
$$= \frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} \rho_I I\ (m - 1)$$
$$= \frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} \rho_i\ (m - 1) \qquad (4.9)$$

where ρ_i is equal to $\rho_I I$.

The total variance (4.7) can now be written as

$$\frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} \{1 + \rho_s(b - 1) + \rho_i(m - 1)\} \qquad (4.10)$$

The design effect becomes

$$\text{Deff} = 1 + \rho_s(b-1)$$

and by analogy, the interviewer effect is

$$\text{Inteff} = 1 + \rho_i(m-1)\,.$$

The design factor is

$$\text{Deft} = \sqrt{\text{Deff}}$$

and the interviewer factor is

$$\text{Inteft} = \sqrt{\text{Inteff}}\,.$$

4.4 Partitioning the Total Variance in Lesotho and Peru

In this section two variables subject to considerable interviewer effect (see table 4.1L) are considered in detail - *First birth interval* and *Ever-use of contraception.* For each the total variance is presented in terms of its four components: simple sampling variance, simple response variance, correlated sampling variance and correlated response variance (interviewer variance). In order to put the results into perspective, two additional variables are also considered in section 4.5 - *Age at marriage* and *Children ever born* - neither of which shows any evidence of interviewer effect.

4.4.1 Partitioning the total variance - *first birth interval*

Table 4.2L presents the results for *First birth interval.* The first column gives the sizes of the base variance components - $\hat{\sigma}_y^2$, $\hat{\sigma}_\varepsilon^2$, *roh* $\hat{\sigma}_y^2$ and $\rho_I\hat{\sigma}_\varepsilon^2$ (note that these components could not be estimated were it not for *both* the interpenetration *and* the reenumeration). The next three columns give the total variance for three special cases: (1) the actual design with

average interviewer workload (m) of 100 and average cluster take of 36.8; (2) a hypothetical design with average interviewer workload of 1 with cluster take unchanged; (3) a simple random sample (ie $b = 1$) with $m = 1$.

Table 4.2L Total variance: first birth interval (Lesotho)

	Base components	$m = 100$ $b = 36.8$	$m = 1$ $b = 36.8$	$m = 1$ $b = 1$
Correlated response variance (CRV)	25.46	2520.5	-	-
Correlated sampling variance (CSV)	6.79	243.0	243.0	-
Simple response variance (SRV)	330.80	330.8	330.8	330.8
Simple sampling variance (SSV)	234.90	234.9	234.9	234.9
Total variance (TV)		3329.2	808.7	565.7

Figure 4.3L presents a diagrammatic representation of the results in table 4.2L. Column I gives the estimate of the actual total variance and its components for the survey design used in Lesotho. The magnitude of the correlated sampling and response variances are based on the values of b and m actually used, ie $b = 36.8$ and $m = 100$. It is clear from column I that the total variance is dominated by the interviewer variance, which accounts for some 76 per cent of the total variance. The correlated sampling variance accounts for 7 per cent, the simple response variance for 10 per cent and the simple sampling variance for 7 per cent.

Figure 4.3L here

Column VI shows only the simple total variance. This is the estimate of the total variance that would be obtained if s^2/n were used as the estimator, ie if the variance were estimated as though for a simple random sample. In this case the total variance would be underestimated by a factor of more than five.

Column V shows the quantity actually estimated in practice for WFS surveys. This is the

estimate provided by using the correct formula for the sampling variance plus the simple response variance. The only component of the total variance neglected by this estimate is the correlated response variance. In this case the total variance would be underestimated by a factor of four.

Columns II, III and IV give an indication of the way in which the total variance could be reduced by changing the field strategy, within a fixed total sample size. Column II gives the total variance for a design in which the number of interviewers is doubled, keeping the sample size unchanged. The effect is to cut by half the contribution of interviewer variance to the total variance, due to the reduction of the interviewer workload m and consequently of the term $\rho_I\sigma_\varepsilon^2(m - 1)/n$. It is assumed in this case that the quality of the interviewers is not affected by increasing their number. Columns III and IV indicate the effect of reducing the interviewer workload to 30 and 10 respectively under the same assumption. In principle, column V is the variance obtained when $m = 1$, ie when each respondent is interviewed by a different interviewer.

The impact of the interviewer variance for *First birth interval* for this design is such that the change in field strategy indicated in column II would lead to a reduction of the actual total variance by 38 per cent. The further increase in the number of interviewers represented by column III would reduce the total variance by a further 15 per cent; an increase to ten times the original number of interviewers would give a total variance equal to less than one-third of the actual total variance.

Column VII is the minimum variance possible for a sample of size n (assuming no stratification). This would be the case if a simple random sample of size n were selected and if the measurement were perfect, ie if there were no response errors of any kind. Column VIII represents the actual total *sampling* variance for the design used. .

The corresponding data for the same variable -- *First birth interval* -- for Peru are given in figure 4.3P. Column I of figure 4.3P represents the estimate of the total variance for the Peru

Fertility Survey. In the main survey the average interviewer workload (m) was 101 and the average number of individuals interviewed in each primary sampling unit (b) was 14.

Figure 4.3P here

The effect of changing the value of m can be seen from columns I to V of figure 4.3P. Column I gives the estimate of the actual total variance and its components for the survey design used in Peru. The magnitude of the correlated sampling and response variances are based on the values of b and m actually used, ie $b = 14$ and $m = 101$. It is clear from column I that the total variance is dominated by the interviewer variance, which accounts for some 61 per cent of the total variance. The correlated sampling variance accounts for 9 per cent, the simple sampling variance for 13 per cent and the simple response variance for 17 per cent.

Column VI shows only the simple total variance. This is the estimate of the total variance that would be obtained if s^2/n were used as the estimator, ie if the variance were estimated as though for a simple random sample. In this case the total variance would be under-estimated by a factor of more than three.

Column V shows the quantity actually estimated in practice for WFS surveys. This is the estimate provided by using the correct formula for the sampling variance. In fact it estimates the total sampling variance plus the simple response variance. The only component of the total variance neglected by this estimate is the correlated response variance.

Columns II, III and IV give an indication of the way in which the total variance could be reduced by changing the field strategy, within a fixed total sample size. Column II gives the total variance for a design in which the number of interviewers is doubled, keeping the sample size unchanged. The effect is to cut by half the contribution of interviewer variance to the total variance, due to the reduction of the interviewer workload m and consequently of the term $\rho_I \ \sigma_\varepsilon^2 \ (m-1)/n$. It is assumed in this case also that the quality of the interviewers is not affected by increasing their number. Columns III and IV indicate the effect of reducing the

interviewer workload to 31 and 11 respectively under the same assumption. In principle column V is the variance obtained when $m = 1$, ie when each respondent is interviewed by a different interviewer.

Column VII is the minimum variance possible for a sample of size n (assuming no stratification). This would be the case if a simple random sample of size n were selected and if the measurement were perfect, ie if there were no response errors of any kind. Column VIII represents the actual total sampling variance for the design used.

Figure 4.4P is an alternative way of looking at the information in columns I to IV. Each bar shows the relative contribution of the four components of variance for one of the six sets of circumstances.

Figure 4.4P about here

4.4.2 Partitioning the total variance - *ever-use of contraception*

The results for *Ever-use of contraception* for Lesotho are given in table 4.3L and in figure 4.5L. The situation is even more dramatic in this case. By comparison with the simple sampling variance and the simple response variance the correlated variance components are overwhelming, and between them they account for more than 90 per cent of the total variance. The difference between column VI and column V highlights the necessity for proper estimation of sampling variance. Ignoring the effect of the clustering in the sample design would lead to an underestimation of almost 60 per cent. The contrast between columns V and I shows that for this variable also the total variance is dominated by the interviewer variance, accounting as it does for almost 78 per cent of the total variance. This situation, of course, is due not only to the intercorrelation between the response deviations but also to the large average workload size. Columns II, III and IV show the effect of reducing the workload size, and demonstrate how this dominance by the interviewer variance can be radically altered. With an average workload size of $m = 10$, for instance, the interviewer variance - other things being equal - would account for less than a quarter of the total variance.

Table 4.3L Total variance: ever-use of contraception in Lesotho

	Base components	$m = 100$ $b = 36.8$	$m = 1$ $b = 36.8$	$m = 1$ $b = 1$
CRV	0.0130	1.2848	-	-
CSV	0.0059	0.2102	0.2102	-
SRV	0.0850	0.0850	0.0850	0.0850
SSV	0.0695	0.0695	0.0695	0.0695
TV	-	1.6495	0.3647	0.1545

Figure 4.5L here

The results for *ever-use of contraception* for Peru are given in figure 4.5P. Here also the situation is dramatic. By comparison with the simple sampling variance and the simple response variance the correlated variance components are overwhelming, and between them they account for more than 90 per cent of the total variance. Ignoring the effect of the clustering in the sample design would lead to an under-estimation by a factor of three. The contrast between columns V and I shows that for this variable also the total variance is dominated by the interviewer variance, accounting as it does for almost 75 per cent of the total variance. This situation of course is due not only to the intercorrelation between the response deviations but also to the large average workload size. Columns II, III and IV show the effect of reducing the workload size. Figure 4.6P presents these results in percentage terms and demonstrates how this dominance by the interviewer variance can be radically altered. With an average workload size of $m = 11$, for instance, the interviewer variance - other things being equal - would account for less than a quarter of the total variance.

Figures 4.5P and 4.6P about here

4.5 Summary Measures of the Variances

The results in section 4.4 are not typical of all variables in the fertility surveys. The two

variables described there are those for which the impact of response variance is greatest. In order to put these results in perspective a set of four variables is considered in this section which includes all types of variables in terms of the relative magnitude of the different components of the total variance.

In the results presented here, Deff and Inteff (and consequently Deft and Inteft) are estimated and their estimates will be noted by deff, inteff, deft and inteft. The choice between using variances and standard errors depends on the purpose for which the results are presented. Tables 4.4L and 4.4P provide both for the four variables concerned. The variables are *Children ever born, Age at marriage, First birth interval* and *Ever-use of contraception.*

Table 4.4L Summary measures of the variances and standard errors for four variables for Lesotho

Variable	1/(1-I)	deff (for $b = 36.8$)	inteff	$\sqrt{1/(1-\hat{I}}$	deft	inteft
Children ever born	1.09	1.14	1.00	1.04	1.07	1.00
Age at marriage	1.44	1.00	1.00	1.20	1.00	1.00
First birth interval	2.41	1.44	4.46	1.55	1.20	2.11
Ever-use of contraception	2.22	2.37	8.32	1.49	1.54	2.88

In order to make the first and fourth columns of the table comparable to the others, 1/(1-I) is presented instead of I. This quantity measures the factor by which the simple sampling variance must be multiplied to give the simple total variance.

The variable least affected overall is *Children ever born.* It has a small component of simple response variance relative to the simple sampling variance; the effect of the clustering of the sample on the variance is slight - an increase of only 14 per cent; and there is no evidence of interviewer effect. Taking the simple sampling variance as a base, the total effect of all the other components is to multiply the variance by a factor of 1.24. If the simple total variance is taken as a base, the multiplying factor is 1.14.

Age at marriage is similarly dominated by the simple sampling variance, although the simple response variance in this case accounts for 30 per cent of the simple total variance. There is no increase in the variance due to the clustering of the sample. In other words, the design effect is equal to 1; *Age at marriage* does not differ systematically across clusters. There is also no evidence of any interviewer variance. The overall ratio of the actual total variance to the simple sampling variance is 1.44.

The two remaining variables are very different. In both cases the simple response variance is a substantial element in the simple total variance. Furthermore the design effect and the interviewer effect are large for both variables. The ratio of the total variance to the simple sampling variance is 11.81 for the *First birth interval* and 21.51 for *Ever-use of contraception*; the ratios of the total variance to the simple total variance are 4.90 and 9.69 respectively.

The corresponding results for Peru are presented in table 4.4P below. The overall picture is strikingly similar to that for Lesotho.

Table 4.4P Summary measures of the variance components and the standard errors for four variables for Peru

Variable	*1/(1-1)*	deff	inteff	$\sqrt{1/(1-}$	deft	inteft
Children ever born	1.02	1.14	1.00	1.01	1.07	1.00
Age at marriage	1.25	1.10	1.00	1.12	1.05	1.00
First birth interval	2.27	1.30	2.99	1.14	1.14	1.73
Ever-use of contraception	1.54	3.39	11.02	1.84	1.84	3.32

Again the variable least affected overall is *Children ever born.* It has a very small component of simple response variance relative to the simple sampling variance; the effect of the clustering of the sample on the variance is slight - an increase of only 14 per cent; and there

is no evidence of interviewer effect. Taking the simple sampling variance as a base, the total effect of all the other components is to multiply the variance by a factor of 1.16. If the simple total variance is taken as a base, the multiplying factor is 1.14.

Age at marriage is similarly dominated by the simple sampling variance, although the simple response variance in this case accounts for 20 per cent of the simple total variance. The effect of the clustering of the sample is to multiply the simple total variance by 1.10. There is no evidence of any interviewer variance. The overall ratio of the actual total variance to the simple sampling variance is 1.375.

The two remaining variables are again very different. In both cases the simple response variance is a substantial element in the simple total variance. Furthermore the design effect and the interviewer effect are large for both variables. The ratio of the total variance to the simple sampling variance is 7.49 for the *First-birth interval* and 20.65 for *Ever-use of contraception*; the ratios of the total variance to the simple total variance are 3.29 and 13.41 respectively.

These ratios are easily calculable from the figures given in tables 4.4L and 4.4P. From (4.10) we have that the total variance is:

$$\frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} \{1 + \rho_s(b - 1) + \rho_i(m - 1)\} \qquad (4.10)$$

This can be written as

$$\frac{\sigma_y^2 + \sigma_\varepsilon^2}{n} \{1 + (deff - 1) + (inteff - 1)\} \qquad (4.11)$$

If d^2 = deff and i^2 = inteff, then the total variance is:

$$\frac{\sigma_y^2 + \sigma_\varepsilon^2}{n}(d^2 + i^2 - 1) \qquad (4.12)$$

The expression (4.12) illustrates the advantage of working directly with the variances and the ratios of variances.

It is more relevant in the context of interval estimation to work with the standard errors. The effect on the standard error of each of the components of variance is given in the second half of tables 4.4L and 4.4P. They cannot however be presented in such simple additive form.

4.6 Confidence Intervals

To illustrate the effects of the variance components on the four variables considered above tables 4.5L and 4.5P give the width of the 95 per cent confidence intervals using the three possible estimation procedures. The sample mean is also given for each variable, and column (1) gives the simple sampling error[1] (ie σ_y^2/n).

Table 4.5L Width of 95 percent confidence interval for four variables using different estimates of the total error (based on n = 609 cases)

Variable	Mean	Simple sampling error (1)	Simple total error (2)	(2)×deft (3)	Correct estimated standard error (4)
Children ever born	3.19	0.162	0.169	0.180	0.180
Age at marriage	17.90	0.173	0.208	0.208	0.208
First birth interval	25.96	1.000	1.550	1.860	3.430
Ever-use of contraception	0.23	0.0172	0.0256	0.0394	0.0797

1 The precise interpretation of the simple sampling error for binary variables is well explicated by Biemer and Stokes in Biemer et al (1989).

Table 4.5P Width of 95 per cent confidence interval for four variables in Peru using different estimates of the total error (based on n=1198 cases)

Variable	Mean	Simple sampling error (1)	Simple total error (2)	(2)×deft (3)	Correct standard error (4)
Children ever born	4.66	0.356	0.360	0.384	0.384
Age at marriage	19.90	0.428	0.476	0.500	0.500
First birth interval	11.30	1.40	2.12	2.400	3.850
Ever-use of contraception	0.56	0.0460	0.0572	0.1052	0.2094

The possible estimates of the standard error are used in columns (2), (3) and (4) in calculating the width of the confidence interval. Column (2) is calculated using s^2/n (where s^2 is the sample variance) as the estimate of the total variance; column (3) uses the appropriate calculation for a complex sample design where the data are free of correlated response variance; and column (4) gives the correct estimate of the total error.

The variable *Children ever born* illustrates the position when neither the correlated sampling variance nor the correlated response variance has much impact. Similarly the various estimates for *Age at marriage* differ little from one another. It should be noted however that for some variables not given in table 4.6 the design effect is important even though there is no interviewer effect.

For the remaining two variables the situation is very different both in Lesotho and Peru.

In Lesotho, for the *First birth interval* the width of the confidence interval using s^2/n to estimate the variance (column 2) would be 1.55; using the standard (correct) estimate of sampling variance (column 3) the width would be 1.86. When the interviewer effect is taken

into account the interval is seen to be 3.43, a *further* increase of 84 per cent. For *Ever-use of contraception* the disparity is even more striking. Column (2) gives a confidence interval of width 0.0256. Once the design effect is introduced, this increases to 0.0394 (column 3), a rise of 54 per cent. The interviewer effect increases the confidence interval to 0.0797, a *further* rise of over 100 per cent.

The effects in Peru are comparable. In Peru for the *First birth interval* the width of the confidence interval estimated using the correct procedure to estimate sampling error would be 2.40. When the interviewer effect is taken into account the width of the confidence interval is seen to be 3.85 - an increase of 60 per cent over the usual estimate. For *Ever-use of contraception* the adjustment is even more striking. Column (3) gives a confidence interval of width 0.1052; the correct confidence interval is of width 0.2094 - an increase of almost 100 per cent.

The importance of both components of correlated variance can also be illustrated by considering the true confidence level for the estimates constructed using columns (2) and (3). Table 4.6L gives the results for Lesotho.

The results are startling and disconcerting. The true confidence level is acceptable for *children ever born* and *age at marriage*. For *First birth interval* the results are little short of appalling. Taking the (relatively sophisticated) estimate from column 3 (which is better than most survey reports would offer) the true confidence level corresponding to the 95% interval is 71%, that corresponding to the 99% level is 84%. In the case of *Ever-use of contraception* the situation is even worse, with true confidence levels of 67% and 80% respectively. Put another way, for *Ever-use of contraception* a significance test at a (nominal) 5% level would actually be being tested at a 33% level.

Table 4.6L Apparent and true confidence levels for confidence intervals constructed using columns (2) and (3) of table 4.5L (Lesotho).

Variable	Estimate	95 % confidence level		99 % confidence level	
		Apparent confidence level (%)	True confidence level (%)	Apparent confidence level (%)	True confidence level (%)
Children ever born	Col. (2)	95	93	99	98
	Col. (3)	95	95	99	99
Age at marriage	Col. (2)	95	95	99	99
	Col. (3)	95	95	99	99
First birth interval	Col. (2)	95	63	99	76
	Col. (3)	95	71	99	84
Ever-use of contraception	Col. (2)	95	47	99	59
	Col. (3)	95	67	99	80

For Peru similar calculations can be carried out. The results are comparable: for the *First birth interval* the apparent 95 per cent confidence interval is actually a 78 per cent confidence interval; for *Ever-use of contraception* the true confidence interval is 68 per cent.

While not diminishing the importance of the results above, it should be borne in mind that the first two variables in table 4.6L are more representative of the generality of variables from WFS surveys than are the last two. Furthermore the figures in table 4.6L are based on estimates of the variance components and these estimates are themselves subject to sampling error. The problem of estimating the variance of the estimates of the components of variance is dealt with in chapter 7.

Chapter 5: FURTHER ISSUES IN INTERVIEWER VARIANCE

The discussion of the implications of interviewer variability has been confined so far to the effect on individual items (variables) for the whole sample. In this chapter three extensions are presented. The first deals with the impact of interviewer variance on summative indexes constructed from a set of variables, as in attitude scales. The second looks at the structure of interviewer effect across a set of variables, attempting to extract from the pattern of effects information about the variables themselves. The third considers the effect of interviewer variance on individual variables for subclasses of the sample.

5.1 Interviewer Variance for Indexes

With attitudinal data, typically, several items (variables) are combined into a scale in order to obtain reliable measures of the underlying dimension. This can be done in a variety of ways ranging from factor analytic methods to simple summated scores; in practice one is concerned with indexes derived from sets of items. In this area therefore it is more important to examine the effect of interviewer variability on these indexes rather than on individual items in isolation. In this section the interviewer effect on the mean of a category of items is considered.

5.1.1 A simple theoretical model

It is more convenient for this presentation to re-state the model from chapter 2 in slightly different terms; the essence of the model - and the estimation process - are unaffected. The correlations among response deviations that generate the interviewer variance between the response deviations may be brought about in a number of ways. In Case II in chapter 2 (section 2.3.2) and the corresponding analysis in chapter 4 (sections 4.2, 4.3 and 4.4) no particular form is assumed for the mechanism that generates the values of ρ_i. In the next two sections the ANOVA model (originally offered in this context by Kish, 1962) is used as it provides an intuitively attractive description of how the intra-interviewer correlations are generated.

The sample of size n is assumed to be divided at random into k independent random samples of size m ($n = mk$). It is further assumed that k interviewers are chosen at random from a large population of interviewers and that one of the subsamples is allocated to each interviewer. Denote each observation by y_{ijt}, where $i = 1, ..., k$ denotes the interviewer, $j = 1, ..., n$ denotes the sample element, $t = 1, ..., T$ denotes the trial (these are the (hypothetical) repetitions of the survey).

In addition to the terms defined already, we need to define

$E(y_{ijt} \mid ij) = y_{ij}$, the expected value obtained by interviewer i for element j (5.1)

Releasing the conditioning on the interviewer, let $E(y_{ijt} \mid j) = y_j$ (5.2)

Thus

$$\begin{aligned} \epsilon_{ijt} &= y_{ijt} - y_j = (y_{ijt} - y_{ij}) + (y_{ij} - y_j) \\ &= e_{ijt} + \alpha_i \end{aligned}$$

$$y_{ijt} = y_j + \alpha_i + e_{ijt} \qquad (5.3)$$

A number of assumptions are built into this model. First, the overall response deviation for each element is split into two additive components α_i and e_{ijt}. The α_i represents the *systematic* effect of interviewer i to push the responses in a particular direction. For example, an interviewer with strong right-wing views might consistently influence the responses in a particular way. Also, it is assumed that the expected value of the e_{ijt} for interviewer i and item j is equal to zero. This follows from (5.1).

For the population of interviewers from which the k interviewers are drawn, equation (5.2) implies that $E(\alpha_i) = 0$, in other words that we are dealing with *compensating* interviewer biases. Any systematic effect which is common to all interviewers is part of the *response bias* as defined in (2.3) in chapter 2.

The model is therefore

$$y_{ijt} = y_j + \alpha_i + e_{ijt} \tag{5.3}$$

where

$$E(\alpha_i) = E(e_{ijt}) = 0$$
$$\text{Var}(\alpha_i) = \sigma_\alpha^2,\ \text{Var}(e_{ijt}) = \sigma_{ei}^2$$

It is further assumed that the α_i and e_{ijt} are uncorrelated, and that both α_i and e_{ijt} are uncorrelated with y_j. Then,

$$\text{Var}(y_{ijt}) = \sigma_y^2 + \sigma_\alpha^2 + \sigma_e^2$$

where

$$\sigma_e^2 = \frac{1}{N}\sum_{i=1}^{N} \sigma_{ei}^2 \quad .$$

The sampling variance is σ_y^2 ; the simple response variance σ_ϵ^2 is equal to $\sigma_\alpha^2 + \sigma_e^2$.

$$\text{Cov}(y_{ijt}, y_{ij't}) = -\frac{\sigma_y^2}{N-1} + \sigma_\alpha^2$$

This is the covariance between observations on two different elements by the same interviewer. The first term is due to sampling without replacement from a finite population

$$\text{Cov}(y_{ijt}, y_{i'j't}) = -\frac{\sigma_y^2}{N-1} \tag{5.4}$$

The second term is due to the correlation between the response deviations due to the interviewer.

The variance of $\bar{y}_t$ can then be written as

$$\text{Var}(\bar{y}_t) = \frac{\sigma_y^2}{n} \cdot \frac{N-n}{N-1} + \frac{\sigma_\alpha^2}{k} + \frac{\sigma_e^2}{n}$$

Since $\sigma_\epsilon^2 = \sigma_\alpha^2 + \sigma_e^2$, this can be written as

$$\mathrm{Var}(\bar{y}_t) = \frac{\sigma_y^2}{n} \cdot \frac{N - n}{N - 1} + \frac{\sigma_\epsilon^2}{n}\left(1 + (m - 1) \cdot \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2}\right)$$

Writing $\rho_1 = \sigma_\alpha^2/\left(\sigma_\alpha^2 + \sigma_e^2\right)$, this gives

$$\mathrm{Var}(\bar{y}_t) = \frac{\sigma_y^2}{n} \cdot \frac{N - n}{N - 1} + \frac{\sigma_\epsilon^2}{n}(1 + (m - 1)\rho$$

This can be re-written, ignoring the finite population correction and conditioning on the simple total variance, as

$$\mathrm{Var}(\bar{y}_t) = \frac{\sigma_y^2 + \sigma_\epsilon^2}{n}(1 + (m - 1)\rho_i) \qquad (5.5)$$

where ρ_i is the usual within interviewer correlation coefficient (see also chapter 4), and is equal to

$$\rho_i = \frac{\sigma_\alpha^2}{\sigma_y^2 + \sigma_\alpha^2 + \sigma_e^2} \quad .$$

The assumptions made in (5.3) simplify the expression considerably but their possible implications should not be overlooked. In particular it is assumed that the $\{y_j \mid j\}$ and the α_i are uncorrelated. This will not necessarily be the case. It is possible that the particular elements included in the sample will influence the interviewer effect. This is part of the extension to the model described in Case III in section 2.3.2 in chapter 2. The analysis above can be applied to correlated errors due to any other part of the measurement or recording process.

For simplicity of presentation we will use ρ without subscript in the following to represent ρ_i. As the data do not involve any re-enumeration, the subscript t will not be included.

Applying this model to a single item in an attitude scale, for each item we have therefore:

$$y_{ij} = y_j + \alpha_i + e_{ij}$$

This can be re-written as:

$$y_{ij} = y_{ij}^{*} + \alpha_i$$

ρ can be written as

$$\rho = \frac{\sigma_\alpha^2}{\sigma_y^2 + \sigma_\alpha^2 + \sigma_e^2} = \frac{V(\alpha_i)}{V(\alpha_i) + V(y_{ij}^{*})}$$

Consider the case where we have a set of items, h=1, ... ,L. The general model is

$$y_{hij} = y_{hij}^{*} + \alpha_{hi} \qquad (h = 1, ..., L;\ i = 1, ..., m;\ j = 1, ..., k)$$

Let

$$z_{ij} = \sum_h^L y_{hij}/L = \sum_h^L y_{hij}^{*}/L + \sum_h^L \alpha_{hi}/L$$

Let ρ_z denote the intra-interviewer correlation coefficient for the variable z. The general expression for ρ_z is

$$\rho_z = \frac{V\left(\sum_h \alpha_{hi}/L\right)}{V\left(\sum_h \alpha_{hi}/L\right) + V\left(\sum_h y_{hij}^{*}/L\right)}$$

$$= \frac{\frac{1}{L^2}\{\sum_h V(\alpha_{hi}) + \sum\sum_{u \neq v} cov(\alpha_{ij}, \alpha_{vi})\}}{\frac{1}{L^2}\{\sum_h V(\alpha_{hi}) + \sum\sum_{u \neq v} cov(\alpha_{ui}, \alpha_{vi}) + \sum_h V(y_{hij}^{*}) + \sum\sum_{u \neq v} cov(y_{uij}^{*}, y_{vih}^{*})\}}$$

This expression does not provide any obvious guide to the relationship between ρ for the individual items and ρ for the mean in the general case. However, by making certain simplifying assumptions, we can gain an insight into what the relationship may be in particular cases, and illustrate these cases from the data. Let $\bar{\rho}$ denote the average of the ρ values for the L items.

Case I Items equally affected by correlated interviewer error

Assume that $V(\alpha_{hi}) = V(\alpha_i)$ for all h, and that $V(y^*_{hij}) = V(y^*_{ij})$ for all h.

The expression for ρ_z reduces to

$$\rho_z = \frac{V(\alpha_i)/L + \sum\sum_{u \neq v} corr(\alpha_{ui}, \alpha_{vi}) \cdot V(\alpha_i)/L^2}{V(\alpha_i)/L + \sum\sum_{u \neq v} corr(\alpha_{ui}, \alpha_{vi}) \cdot V(\alpha_i)/L^2 + V(y^*_{ij})/L + \sum\sum_{u \neq v} corr(y^*_{uij}, y^*_{vij}) \cdot V(y^*_{ij})/L^2}$$

$$= \frac{V(\alpha_i)\left[1 + \frac{\sum\sum corr(\alpha_{ui}, \alpha_{vi}}{L}\right]}{V(\alpha_i)\left[1 + \frac{\sum\sum corr(\alpha_{ui}, \alpha_{vi})}{L}\right] + V(y^*_{ij})\left[1 + \frac{\sum\sum corr(y^*_{uij}, y^*_{vij})}{L}\right]}$$

If we denote the average corr$(\alpha_{ui}, \alpha_{vi})$ by $\bar{r}_\alpha$ and the average of corr (y^*_{uij}, y^*_{vij}) by $\bar{r}_{y*}$, then

$$\rho_z = \frac{V(\alpha_i)[1 + (L - 1)\bar{r}_\alpha]}{V(\alpha_i)[1 + (L - 1)\bar{r}_\alpha] + V(y^*_{ij})[1 + (L - 1)\bar{r}_{y*}]}$$

Therefore

$$\text{if } \bar{r}_\alpha > \bar{r}_{y*}, \quad \rho_z > \bar{\rho},$$
$$\text{if } \bar{r}_\alpha = \bar{r}_{y*}, \quad \rho_z = \bar{\rho},$$
$$\text{and} \quad \text{if } \bar{r}_\alpha < \bar{r}_{y*}, \quad \rho_z < \bar{\rho}.$$

The assumptions made in this case are fairly restrictive. By assuming that $V(\alpha_{hi})$ and $V(y^*_{hij})$ are constant for all h we determine that ρ is fixed for all h, i.e. for all items in the category. However, the result above has some important implications. Firstly, the higher the correlations between the y^*_{hij} - the item scores with the interviewer effect removed - the less

likely it is that ρ_z will exceed $\bar{\rho}$; this is reassuring in the case of attitude scales where items are selected for inclusion on the basis of high internal consistency. Whatever the values of the ρ's, to look at the ρ-values for the items individually may be seriously misleading. If $\bar{r}_\alpha$ is large, then the effect on the mean may be greater than the average effect; in particular if $\bar{r}_\alpha = +1$, than ρ_z will always be greater than or equal to $\bar{\rho}$. Alternatively if the effect is in a different direction for each item i.e. $\bar{r}_\alpha$ is low or negative, then the effect on the mean may be considerably less than the average effect. In particular if $\bar{r}_\alpha = -1/(L-1)$ then there will be no interviewer effect on the mean.

Case II One item affected by correlated interviewer error

An important special case arises when one item shows a much greater interviewer variability than the other items in the category. As the simplest example of this situation, consider the case when interviewer variance is present for one item only.

Therefore, $V(\alpha_{hi}) = 0 \ (h = 1, \ldots, L-1)$, and $V(\alpha_{hi}) = a \ (h = L)$ and hence $\rho = 0 \ (h = 1, \ldots, L-1)$, and $\rho = a/(a + V_L) \ (h = L)$, where $V_L = V(y^*_{Lij})$.

Hence the average ρ for the items in this category is

$$\bar{\rho} = \frac{a/L}{a + V_L} \tag{5.6}$$

The general expression for ρ_z gives

$$\rho_z = \frac{a/L^2}{a/L^2 + \sum_h V(y^*_{hij})/L^2 + \sum\sum_{u \neq v} cov(y^*_{uij}, y^*_{vij})/L^2} \tag{5.7}$$

$$= a/L^2 + V\left[\frac{\sum y^*_{hij}}{L}\right]$$

Comparing (5.6) and (5.7) we find that

$$\rho_z > \bar{\rho} \text{ if } \frac{(L-1)a}{L^2} > V\left[\frac{\sum y^*_{hij}}{L}\right] - \frac{V_L}{L},$$

$$\rho_z = \bar{\rho} \text{ if } \frac{(L-1)a}{L^2} = V\left[\frac{\sum y^*_{hij}}{L}\right] - \frac{V_L}{L},$$

and

$$\rho_z < \bar{\rho} \text{ if } \frac{(L-1)a}{L^2} < V\left[\frac{\sum y^*_{hij}}{L}\right] - \frac{V_L}{L}.$$

The condition above is complex but some general remarks can be made. Firstly, the larger the value of a and the larger V_L is in relation to the variances of the other items, the more likely it is that $\rho_z > \bar{\rho}$. Secondly, the higher the correlations between the y^*_{hij}, the less likely it is that $\rho_z > \bar{\rho}$.

5.1.2 Results from the absenteeism pilot survey

The details of the study are given in section 3.1 of chapter 3. In this investigation, the average score for each individual was calculated for each of the job satisfaction categories, and the $\hat{\rho}$s for the category means were calculated. Table 5.1 presents the results and also includes the average of the $\hat{\rho}$s for the individual items included in the category mean. Of the twenty cases, eleven yielded negative $\hat{\rho}$s for the category means whereas only five negative values were obtained for the averages of the $\hat{\rho}$s. This is not altogether surprising since the range of negative values is restricted to 0 to $-1/(k-1)$ (Kish, 1962), while the positive values can range from 0 to +1. Also, under the assumed model, negative values of $\hat{\rho}$ arise only as a result of sampling variation and averaging the $\hat{\rho}$s will tend to reduce the sampling errors. In no case was there a positive value for $\hat{\rho}_z$ and a negative value for $\bar{\rho}$. Nevertheless it would be wrong to conclude that the $\bar{\rho}$ overestimates $\hat{\rho}_z$. There are examples even in this small set of data which are similar to each of the special cases derived from the theoretical model.

Table 5.1 Interviewer effect for category means: absenteeism data

Category	No. of items	Range of indiv. $\hat{\rho}$s	Average $\hat{\rho}$ $(\bar{\rho})$	$\hat{\rho}$ for category mean $(\hat{\rho}_z)$
1 Work group	7	-0.04 to +0.12	+0.03	-0.04
2 Innovation	8	-0.07 to +0.03	-0.01	-0.02
3 Responsibility	6	-0.07 to +0.30	+0.06	+0.27
4 Pay and promotion	6	-0.05 to +0.07	+0.03	-0.02
5 Identification	6	-0.06 to +0.10	+0.01	-0.05
6 Supervision	10	-0.07 to +0.06	+0.02	+0.06
7 Status	6	-0.03 to +0.04	+0.01	+0.05
8 Change	4	-0.04 to +0.09	+0.03	-0.03
9 Management	5	-0.04 to +0.10	+0.04	+0.10
10 Decision making	6	-0.07 to +0.08	-0.01	-0.03
11 Work group	7	-0.08 to +0.16	+0.03	-0.02
12 Innovation	7	-0.03 to +0.14	+0.04	-0.02
13 Responsibility	5	-0.05 to +0.02	-0.02	-0.07
14 Pay and promotion	7	-0.04 to +0.26	+0.08	+0.17
15 Identification	6	-0.07 to +0.02	-0.03	-0.04
16 Supervision	13	-0.05 to +0.09	+0.03	+0.02
17 Status	4	-0.07 to +0.17	+0.03	+0.03
18 Change	4	-0.01 to +0.02	+0.01	-0.06
19 Management	5	-0.04 to +0.11	+0.04	+0.07
20 Decision making	6	-0.06 to -0.04	-0.05	-0.08

In categories 7, 18 and 20 the $\hat{\rho}$s for the individual items are approximately equal. These categories correspond to case I in the model. In category 7 $\hat{\rho}_z$ is considerably greater than $\bar{\rho}$, whereas in categories 18 and 20 $\hat{\rho}_z$ is less than $\bar{\rho}$. Approximating to case II in the model we have categories 1, 2, 3, 12, 14 and 17. In each category one value of $\hat{\rho}$ is considerably greater than the others. However the relationship between $\hat{\rho}_z$ and $\bar{\rho}$ varies. For categories 3 and 14 $\hat{\rho}_z > \bar{\rho}$ and is quite close to the single largest value of $\hat{\rho}$. For categories 1 and 12 $\hat{\rho}_z < \bar{\rho}$ and is quite close to the smallest single value of $\bar{\rho}$. For categories 2 and 17 $\rho_z = \bar{\rho}$.

The data from this investigation illustrate that the examination of the $\hat{\rho}$ values for the individual items may not give a good guide to the effect of interviewer variability if the analysis of the survey results is to be done in terms of sample statistics derived from the observations. The results do not provide definitive conclusions about the effect of interviewer variability on category means, but the cases above do demonstrate that the use

of interviewer effect on individual variables as a criterion is not sufficient.

5.1.3 Results from the noise annoyance survey

The noise annoyance survey is described in section 3.2 of chapter 3; the data from the study support the conclusions above. Twenty-nine (29) of the attitude items formed the General Health Questionnaire (GHQ) and a further ten (10) formed a noise annoyance score. The individual ρ-values - separating the two scales - are given in figure 3.1 in chapter 3. From the figure it can be seen that the individual items in the annoyance scale are, in general, more sensitive to interviewer effect (median ρ=0.031) than those in the GHQ (median ρ=0.005). In each case however the primary function of the items is to form components of an additive scale, the total score for which is the variable (or measure) of interest to the researcher. Therefore we computed directly the ρ-values for the scale scores for each of the two scales. These, together with the estimate of the corresponding multiplier effect, are given in Table 5.2.

Table 5.2 Values of ρ and the variance multiplier for the GHQ and annoyance scale

Scale	$\hat{\rho}$	m	$1+\hat{\rho}(m-1)$
GHQ total score	0.06	29.5	2.77
Annoyance score	0.02	29.5	1.59

The results in Table 5.2 are illuminating. Although the individual items in the GHQ scale are less sensitive to interviewer effect than those in the annoyance scale, the scale score is considerably more sensitive for the GHQ. On this evidence the usual estimate of variance underestimates the true variance by 64%. Even for the less sensitive annoyance scale the underestimation is of the order of 37%. The reason for the greater sensitivity of the GHQ scale is that the direction of the net distortion of the responses by the interviewer is similar for the items in the scale, whereas for the annoyance scale different items are affected in different ways.

This analysis suggests two principal conclusions. First, at the developmental (pilot) stage of a survey it would be appropriate to incorporate a similar design of the interviewer allocation so that items sensitive to interviewer effect can be rejected (or modified) before the main field work is carried out. Secondly, more work needs to be done on the relationship between the ρ-values for individual items and the ρ-value obtained on the overall scale score. As the preceding analysis has shown, in the presence of interviewer effect the correlation between the observed y's can misrepresent the correlation between the underlying y's. The correlation can be distorted in either direction depending on the relationship between the interviewer impacts (the α's) for the items making up the scale. If the α's are more strongly correlated than the y's the correlations be exaggerated, if the α's are less correlated than the y's the correlations will be attenuated. As a possible precaution where interviewer effect is not estimated in a pilot survey, a single interviewer should carry out all the interviewing; this would avoid mis-estimation of ρ_{yy} in cases where the α's were correlated across interviewers and consequent errors in the choice of items for the scale.

5.2 Structure of Interviewer Effect

In this section an attempt is made to ascertain the underlying structure of the effect which interviewers have on items in an attitude scale. The work described in this section arose first from dissatisfaction with the existing method of analysing interviewer effects i.e. univariate analysis of variance. Most attitudinal questionnaires consist of items designed to measure a small number of basic dimensions. The analysis of the data is concerned largely with identifying these dimensions and their relationship to one another and to the dependent variable or variables. It seemed reasonable therefore that in examining interviewer effects, the analysis should also deal with the items as a multivariate set. The second problem which prompted this investigation was the conflicting evidence which has emerged from other studies in which an attempt was made to find the factors which affected the magnitude of the interviewer effects. Thirdly it seemed that, instead of viewing the existence of interviewer effect as a necessary evil, it might be possible to use the interaction between the interviewer and respondent to provide information about the subject matter rather than to obscure it.

5.2.1 Absenteeism survey

Multivariate analysis of variance (test for the presence of multivariate interviewer effect)

Here we look again at the data from the absenteeism survey. The methods of analysis used in univariate studies gives some guidelines as to the type of multivariate analysis which might be profitable. The earlier analysis of interviewer effects used the ratio of σ^2_α to σ^2_y as an indication of the importance of the interviewer effect for a particular item. If we consider a set of L items $(y_1, y_2, ..., y_L)$ we have for each interviewer a vector of the mean values of these variables obtained for the respondents whom s/he interviewed. Thus, we have k vectors $(\bar{y}_{i1}, \bar{y}_{i2}, ..., \bar{y}_{iL}, i = 1, ..., k)$, and we wish to test the null hypothesis that these arose from the same L-variate normal distribution. A test of this hypothesis is provided by Wilks' Lambda, which is the ratio of the determinant of the variance - covariance matrix of the interviewer effects to the determinant of the variance - covariance matrix of the residuals. An approximate F test is given by Rao. This test is analogous to the F test in the univariate analysis of variance.

This approach is reasonable in the case of the pilot study data where k=5 and the respondents are randomised among all the interviewers. The results of the test are given for the data from the pilot study are given in table 5.3. Thirty three (33) items dealing with job satisfaction were included in the analysis. With five interviewers we obtain four roots of the determinant of the variance-covariance matrix of the effects. We first test all four roots, then eliminate the largest and test the remaining three, and so on.

Table 5.3: Multivariate analysis of variance: test of significance using Wilks' lambda.

Test of	F	Probability
All four roots	1.444	0.020
Roots 2-4	1.222	0.170
Roots 3-4	0.977	0.536
Root 4	0.844	0.677

On the basis of this test we reject the null hypothesis. It is interesting to note that once the largest root is eliminated the remaining three tested together are not significant. This could indicate that the interviewer effects are in some sense unidimensional; or perhaps two-dimensional, since once we eliminate the two largest roots, all the variation can be explained by chance (P=0.536).

The model for the main study is much more complicated (see section 3.1.1 of chapter 3). As a first stage however it is possible to treat each case separately and test the difference between the mean vectors of the two interviewers involved. For the data from the main study, as mentioned in that section, the test provided significant evidence of the existence of interviewer effecting twenty two of the twenty four cases tested.

Principal component analysis (Examination of Structure of Effects)

Pilot Absenteeism Study

From the analysis of variance the interviewer effects $\{\alpha_{hi}(= 1,..., L;\ i=1,...,m)\}$ can be estimated.

Consider the *Lxm* matrix of effects $|A| = \begin{vmatrix} \alpha_{11} & \alpha_{1m} \\ \alpha_{L1} & \alpha_{Lm} \end{vmatrix}$. In order to examine the structure of

the variation in this matrix principal component analysis may be carried out. The maximum number of components is *(m-1)*. It is possible that a smaller number of components may explain the variation completely. If a single component were to explain all the variation this would mean that each row of the matrix A was a scalar multiple of each other row; or in other words, that there exists a row vector of effects $(\alpha_1\ \alpha_2\ \alpha_3\ \alpha_4\ \alpha_5)$, and each row of the matrix A is a scalar multiple of this vector.

With the data from the pilot study, the first principal component explained 53% of the total variance; the second explained 32%; the third 13% and the fourth only 2%. By examining the correlations between the item interviewer effects and the components, it was possible to identify three dimensions - each closely related to one of the first three components - which explained almost all the variation in the matrix of effects. The sets of items which make up these dimensions are also given in Appendix 2 of O'Muircheartaigh (1974).

Main absenteeism study

Some difficulties arise when attempting to apply the same approach to the data from the main study. For each case (i.e. each firm, or pair of firms, as appropriate) a number of statistics were obtained: (i) the value of $\hat{\rho}$ for each item; (ii) the overall means; (iii) mean deviations; (iv) interviewer means; (v) the difference between the two interviewer means in each case; (vi) the absolute value of the difference between the two interviewer means. Each of these statistics estimates in some sense the `interviewer' effect' for a pair of interviewers for an item. Consequently, it was possible to set up a matrix of `interviewer effects' for the full set of interviewers for each of these estimators. In this chapter, the analysis was concentrated on (i), (v) and (vi) above.

A principal component analysis was carried out on the 8×72 effects matrix in each case. Table 5.4 gives the results. For cases (iv) and (v) four components explained 80% of the variance; for case (vi), five components were necessary.

Table 5.4 Results of principal components analysis on attitudinal variables

Measure		No. of components	No. accounting for 80% of variance
(i)	Estimates of $\hat{\rho}$	7	4 (78.6%)
(v)	Difference in interviewer means	7	4 (78.9%)
(vi)	Absolute difference in interviewer means	7	5 (87.2%)

Using the same procedure as that used in the analysis of the pilot study, sets of items were obtained which were closely related to each of the first three components for each measure. When, in turn, a principal component analysis was carried out on the effects matrix for each of the sets separately, the results in table 5.5 were obtained.

Table 5.5 Principal component analysis of three sets of items for the three measures of interviewer effects

	Measure		
Subset	Difference between interviewer means	$\hat{\rho}$	Absolute difference between interviewer means
1	1st component: 68%	73%	65%
2	1st component: 68%	72%	64%
3	1st component: 69%	81%	77%

The items which made up each subset were not equivalent for the three measures. However, by examining the component scores, it is possible to compare the structure of each of the components in terms of the relationship between the effects of the eight pairs of interviewers. Table 5.6 presents the results.

Table 5.6 Structure of the dimensions of variation for the three measures

Measure	1st dimension	2nd dimension	3rd dimension
Difference between interviewer means	Pairs 6 and 8 vs 3 and 4	Pairs 8 and 5 vs 6 and 4	Pair 7 vs 3 and 6
$\hat{\rho}$	Pairs 6 and 8 vs 3, 4 and 5	Pairs 5 and 8 vs 3	Pair 8 vs 5
Absolute difference between interviewers means	Pairs 6 and 8 vs 1, 3 and 4	Pairs 6 vs 8, 3 and 4	Pair 7 vs 3 and 6

The similarity in structure is reassuring although the differences between the sets of items involved is a little disturbing. As an example, the items for the first three dimensions in the first case above are given below.

Set 1

1. The people I work with would stick up for me if I had trouble with management or my supervisor.
2. I like this factory better than any other place in which I have ever worked.
3. The work I do is important to the company.
4. If I were to begin working again, but in the same occupation as I'm in now, I would be very likely to choose this firm as a place to work.
5. I am proud of the products of this firm.
6. When you complain to your supervisor, how likely is he/she to do something about your complaint?
7. I am consulted when decisions are made concerning my work.
8. The management does not treat you as a person only as a number.
9. Supervision is more strict here than it is in other firms.
10. Supervisors here try to push the workers around.
11. The supervisor has his/her favourites.
12. The supervisors here think they are better than the workers.
13. Some people are very involved in their job. For other people their job is something they have to do. How involved do you feel in your job?

14. How often do you do some extra work or put in some extra effort for your job which is not really required of you?
15. How often do the following changes introduced by management lead to better ways of doing things? (b) Changes in Personnel.
16. I would not like my job to be changed in anyway.
    (Note: refers to change in machinery or method.)
17. I would not like to be transferred to another department.
18. People who work here are given responsibility in their work.

Set 2

1. Does your supervisor take much notice of what you are saying?
2. You don't get any thanks from the supervisor for working hard.
3. With this company, I don't have to worry about losing my job.
4. People work harder in this firm than they do in firms similar to it.
5. Having new people come into the department disrupts things.
6. In my experience here, changes lead to trouble of one kind or another.
7. I like when my work is bit difficult because it is more of a challenge.
8. I feel that my work is a challenge to me.

Set 3

1. If a friend asked me I would tell him/her to try and get a job here.
2. I get annoyed if I hear people criticizing this firm.
3. Which of these statements is closest to the way you feel:
   A. People should not stay out from work unless they are genuinely sick.
   B. If somebody feels he/she needs a day off, he/she is entitled to take it.
4. Would you say you work harder, less hard or about the same as other people doing your type of work? (b) In this firm.

5.2.2 Noise annoyance survey

For the annoyance scale that showed high interviewer effects for individual items, but relatively low interviewer effect for the scale score, the same analysis was carried out. The

univariate analysis of variance presented in section 3.2 of chapter 3 provides information on the sensitivity of individual items to interviewer effect. We would, however, also like to know more about the way in which different subsets of items are affected by the interviewers. In this section we consider the responses to the annoyance scale as a multivariate data set by first estimating the net interviewer biases - the set of α_is in the model given in section 5.1 - for each item and constructing a *matrix of interviewer effects*. A preliminary test of the multivariate item set, using multivariate analysis of variance, indicated that there is a strong interviewer variance component in the data. Furthermore, the analysis suggested that this interviewer effect was not unidimensional. This would indicate that different items in the scale are affected in different ways by the interviewers, or, alternatively, that different interviewers contribute most of the interviewer effect for different items.

To investigate this issue we carried out a principal component analysis of the 8×10 matrix of interviewer effects, i.e. the set of 8 α-values from each of the 10 items on the scale. Table 5.7 gives the results. The first three components describe almost all of the variation in the interviewer effects for the 10-item scale (82.1%). Table 5.8 presents the component score for each interviewer on the 3 main components identified in table 5.7.

Table 5.7 Principal component analyses on matrix of interviewer effect for annoyance scale (8×10)

Component	Eigenvalue	Variance contribution	Cumulated
1	5.02	50.3	50.3
2	1.87	18.7	69.0
3	1.31	13.1	82.1
4	0.97	9.7	91.7
5	0.37	3.7	95.5
6	0.29	2.9	98.4
7	0.16	1.6	100.0

Table 5.8 Component scores for the first 3 principal components

	Component		
Interviewer	1	2	3
1	-0.30	1.07	-1.39
2	0.51	-1.47	0.92
3	0.14	-0.58	-0.26
4	1.23	-0.46	-0.20
5	-1.72	-0.75	-0.34
6	-1.03	0.23	0.48
7	1.03	0.44	-0.90
8	0.13	1.52	1.71

If we consider component 1, for example, we see that most of the variation which it describes is attributable to opposite net biases of the two pairs of interviewers (4 and 7) and (5 and 6). In other words, the interviewer variance results mainly from the fact that two of the interviewers tend to distort the responses substantially in one direction on average, while two others distort the responses in the opposite direction. By direct inspection of the $\{\alpha_i\}$, or by calculating the correlation between the component scores and the $\{\alpha_i\}$, we can identify subsets of items which are affected similarly by the interviewers.

For the 10 items considered here we identified four 'clusters' of items which exhibited different patterns of variation in the $\{\alpha_i\}$. These are presented in table 5.9. The clustering of items is both interesting and interpretable: the items in the first set deal largely with factors associated with specific passive/leisure activities, whereas the fourth set deals with two 'catch-all' items of interference. It is difficult to say why the item 'startle you' stands alone, unless it provoked problems of understanding and delivery for certain respondents and interviewers. Interestingly, items 1 and 9 relate to how 'bothered' the respondent is about aircraft noise nuisance and 'feeling tense and edgy' in general rather than in terms of interference with specific activities. Obviously, replicative studies would be useful, particularly if more data could be obtained as to the behavioural cues used by interviewers.

Table 5.9 Clusters of items corresponding to the components identified in table 5.8

Cluster	Items	Correspond to components (range of correlations)
1	Does aircraft noise ever ...	
	interfere with listening to radio or TV	I
	make the TV picture flicker	(0.75 - 0.95)
	make the whole house vibrate	
	interfere with conversation	
	wake you up?	
2	Does aircraft noise ever ...	
	startle you?	II
		(0.93)
3	Altogether how much are you 'bothered' (by aircraft noise)?	
	Does aircraft noise ever make you feel tense or edgy?	Mixture of I and II
4	Does aircraft noise ever interfere with any activity?	
	Does aircraft noise ever bother, annoy or disturb you in any other way?	III (0.61, 0.89)

These findings seem to confirm the results generated by the absenteeism studies (see also O'Muircheartaigh, 1977) showing that different interviewers behave differently for different sets of items on a scale. The results have possible implications for interviewer briefing and supervision and may help to suggest ways in which interviewer instructions for handling attitudinal items may be improved in the context of a specific survey.

More work needs to be done on the relationship between the ρ-values for individual items and the ρ-value obtained on the overall scale score. Here we have attempted to identify the dimensions on which particular interviewers show substantial deviations from one another for the annoyance scale. The results indicate that, even with the small group of interviewers used in the study, the sets of items which show similar effects form interesting and

interpretable clusters. Combining such differently affected items in the same scale attenuates considerably the impact of interviewer variance.

5.2.3 Application to ordering of variables

There is some evidence in the literature as to how different types of items are affected to a different extent by interviewers. Gales and Kendall found that ambiguity in a question leads to high variability; Hanson and Marks found that contributory factors were (i) interviewer "resistance" to a question, (ii) relatively high ambiguity, (iii) the extent to which additional questioning (probing) tends to alter initial respondent replies; Kish examined three categories of items - critical, ambiguous and other - but found no clear pattern. This problem of categorisation may well stem from the fact that in all these studies an attempt was made to order all the items in a single list. If, however, as this investigation shows, the pattern of variability between interviewers differs for different sets of questions, it is not surprising that the attempts at categorisation have not been completely successful. It may be that if the appropriate dimensions were ascertained, the ordering of the items within each dimension in terms of the magnitude of the interviewer effects might be more readily interpretable.

5.2.4 Summary and Conclusions

The analysis consisted of three stages. In the first stage the full data matrix was analysed using multivariate analysis of variance in order to test for the presence of significant interviewer effects when the data is considered as a multivariate set. In order to investigate the structure further the interviewer effects $\{\alpha_{hi}\}$ were estimated from the analysis of variance for each item.

The second stage consisted of a principal component analysis of the matrix $[A]$ of the interviewer effects. An examination of the correlations between the interviewer effects for each item and each of the principal components showed that the items could be broken down into sets of related items.

The analysis shows that it is possible to consider an attitudinal questionnaire of this kind as

a multivariate set of items, and that it is possible to analyse the data set while taking into account the relationship between the variables. The analysis of attitudinal data is concerned with the relationships between variables, and it is important to identify the pattern of interviewer effects for different variables. By identifying sets of items in which the pattern is similar, it may be possible to interpret the factors in the items from which this similarity is derived. In other words it may be possible to utilise the interviewer effects in order to achieve some appreciation of the subject matter. Finally the division of the complete sets of items into homogeneous groups enables us to interpret more readily the characteristic of the items which determine the sensitivity of the items to interviewer effect.

There is one other use to which an analysis of this kind may be put. As well as deriving information about the items it should be possible to ascertain the characteristics of the interviewers which influence the pattern and magnitude of interviewer effects. For a study of this kind a larger number of interviewers would be necessary and information - both objective and attitudinal - about the interviewers would have to be collected at the time of the investigation. A larger number of interviewers would also make it possible to refine the examination of the dimensions on which interviewer effects vary.

5.3 Interviewer Variance for Subclasses

In common with many other surveys, one of the main objectives of the WFS is to produce separate estimates for subgroups or subclasses of the study population, such as particular demographic, socio-economic or geographic categories. While the number of substantive variables involved may not be very large, the subclasses of interest tend to be much more numerous; each cell of the multiway cross-tabulations of the survey results forms a subclass. Further, much of the analysis of survey results may take the form of comparing and contrasting estimates for different subclasses, resulting in an even larger number of *subclass differences* of interest.

In practice it therefore becomes necessary to confine computation of variances to a selection of subclasses and subclass differences. This approach was used in Verma, Scott and O'Muircheartaigh (1980) in the presentation and analysis of sampling errors for the WFS.

In that paper three groups of subclasses were used: (i) subclasses defined in terms of *demographic* characteristics (age, marriage duration, etc); (ii) subclasses defined in terms of *socio-economic* characteristics (woman's literacy, husband's level of education, occupation, etc); and (iii) a small number of *geographic* subclasses (regional and urbanization classes, for instance). These different subclasses correspond to the major categories by which WFS surveys are cross-tabulated.

Subclasses in the three groups tend to differ in the way in which the elements in the subclasses are distributed across the primary sampling units in the sample. Demographic subclasses are generally fairly uniformly distributed across clusters and form what may be called *crossclasses*. Socio-economic subclasses have a less uniform spread; higher educational groups and non-farming occupations tend to be concentrated in urban areas, for example. These may be called *mixed classes*. By contrast, geographic subclasses are in most cases completely *segregated* - either all or none of the elements in a sample cluster will belong to a subclass. This terminology is due to Kish, Groves and Krotki (1976).

For several purposes it is useful to investigate the relationship between the total variance for an estimator based on the whole sample and the total variance for subclasses and subclass differences: (a) to extrapolate results computed for a particular set of subclasses to numerous other subclasses of interest; (b) to simplify the presentation of results; and (c) to seek stable relationships between the total variance for the whole sample and the total variance for subclasses of particular kinds. In this context, if a stable pattern is found for the relationship, this may provide a better procedure for estimating the total variance for a subclass than direct computation, since each individual estimation is itself subject to a (possibly) large sampling variance.

Three models have been used in the past for the relationship between the variance for the whole sample and the variance for a subclass. The work in this areas has been done for sampling variance only and is described in Kish *et al* (1976) and Verma *et al* (1980). The empirical results obtained have suggested that for crossclasses the intracluster correlation coefficient is approximately stable, although it may increase slightly as the relative size of the crossclass decreases.

In this section that analysis is extended to the more complex case of the total variance. The data available do not permit empirical testing of the model described. The purpose of this section is to investigate the implications of a simple approximate model for the total variance for a crossclass. The algebraic presentation is illustrated by applying it to the total variance found in Lesotho and Peru for the four variables discussed in section 4.5 of chapter 4.

5.3.1 A model

The total variance of the sample mean is:

$$V(\bar{y}) = \frac{\left(\sigma_y^2 + \sigma_\epsilon^2\right)}{n} \{1 + \rho_s(b - 1) + \rho_i(m - 1)\} \tag{5.8}$$

The model proposed here makes a number of assumptions. In particular σ_y^2, σ_ϵ^2, ρ_s and ρ_i are assumed to be the same for the total sample and for the crossclasses. Denoting characteristics of the total sample by the subscript t and those for a crossclass by the subscript c, we therefore have:

$$V_t(\bar{y}) = \frac{\sigma_y^2 + \sigma_\epsilon^2}{n_t} \{1 + \rho_s(b_t - 1) + \rho_i(m_t - 1)\}$$

and (5.9)

$$V_c(\bar{y}) = \frac{\sigma_y^2 + \sigma_\epsilon^2}{n_c} \{1 + \rho_s(b_c - 1) + \rho_i(m_c - 1)\}$$

For a subclass of size $n_c = M_c n_t$ (ie using M_c to denote n_c/n_t

$$\frac{V_c}{V_t} = \frac{\{1 - \rho_s - \rho_i\} + M_c \{\rho_s b_t + \rho_i m_t\}}{M_c \{1 - \rho_s - \rho_i\} + M_c \{\rho_s b_t + \rho_i m_t\}}$$

$$> 1 \text{ if } M_c < 1 \text{ and } 1 - \rho_s - \rho_i \geq 0 \tag{5.10}$$

A limiting case of some interest is that of a simple random sample with no correlated response variance, ie $\rho_s = \rho_i = 0$. In this case

$$\frac{V_c}{V_t} = \frac{1}{M_c} \tag{5.11}$$

In general, (5.10) can be written as:

$$\frac{V_c}{V_t} = 1 + \frac{1 - M_c}{M_c} \cdot \frac{k_1}{k_1 + k_2} \tag{5.12}$$

where $k_1 = 1 - \rho_s - \rho_i$

and $k_2 = \rho_s b_t + \rho_i m_t$

Thus we see that the relationship between the total variance for the whole sample and the total variance for a crossclass can be reduced to a very simple form. The quantity $(1-M_c)/M_c$ is fixed for a subclass making up the proportion M_c of the whole sample. The only quantity which needs to be calculated is $k_1/(k_1 + k_2)$ where k_1 is a simple function of ρ_s and ρ_i, and k_2 takes into account also the average cluster take in the whole sample (b_t) and the average workload size for the whole sample (m_t).

5.3.2 Applications

The model is applied below to the four variables. The derivation shows that the important factors are ρ_s, ρ_i and $k_1/(k_1 + k_2)$. Tables 5.10L and 5.10P give the values of these parameters for the four variables.

Table 5.10L Values of ρ_s, ρ_i and $k_1/(k_1+k_2)$ for Lesotho

Variable	ρ_{cl}	ρ_{int}	$k_1/(k_1+k_2)$
Children ever born	0.004	0.000	0.8712
Age at marriage	0.000	0.000	1.0000
First birth interval	0.012	0.045	0.1908
Ever-use of contraception	0.038	0.084	0.0916

Table 5.10P Values of ρ_s, ρ_i and $k_1/(k_1 + k_2)$ for Peru

Variable	ρ_s	ρ_i	$k_1/(k_1 + k_2)$
Children ever born	0.0113	0.00	0.8638
Age at marriage	0.0078	0.00	0.9021
First birth interval	0.0234	0.02	0.2899
Ever-use of contraception	0.1875	0.10	0.0532

The first two variables are examples of the simple case when there is no interviewer variance and the correlated sampling variance is also relatively small (in fact for one variable in Lesotho both correlated components are zero). The effect of this is to give values of $k_1/(k_1 + k_2)$ close to 1, which is the limiting value for the situation where there is no correlated variance. The third variable is an intermediate case where both components of correlated variance are present and non-negligible. The last variable is an extreme case where the data are subject to large correlated sampling variance and large correlated response variance. The effect of this is seen in the extremely small value of $k_1/(k_1 + k_2)$ - the absolute minimum value for this factor is zero.

The implications of the parameters in tables 5.10L and 5.10P can be seen from tables 5.11L and 5.11P, which give the relative magnitude of V_t and V_c - the values of V_c/V_t are presented for three different subclass sizes. The subclass sizes chosen are $M_c = 0.5$, 0.3 and 0.1. The first corresponds to a subclass which makes up half of the sample, the second to a subclass comprising 30 per cent of the sample, and the third component one tenth of the sample. Most subclasses used in practice fall in this range, although for multiway classifications even smaller subclasses may be involved.

In evaluating the numbers in tables 5.11L and 5.11P it is important to remember that the ratio V_c/V_t must be between $1/M_c$ and 1, where the value $1/M_c$ corresponds to the case where there is no correlated variance, and the total variance is inversely proportional to sample size. For reference, the last row of the table gives the values of $1/M_c$.

Table 5.11L Relative magnitude of V_c and V_t (values of V_c/V_t) for Lesotho

Variable	Cross-class size (M_c)		
	0.5	0.3	0.1
Children ever born	1.87	3.03	8.84
Age at marriage	2.00	3.33	10.00
First birth interval	1.19	1.45	2.72
Ever-use of contraception	1.09	1.21	1.82
No correlated variance	2.00	3.33	10.00

Table 5.11P Relative magnitude of V_c and V_t (values of V_c/V_t) for Peru

Variable	Subclass size (M_c)		
	0.5	0.3	0.1
Children ever born	1.86	3.02	8.77
Age at marriage	1.90	3.10	9.12
First birth interval	1.29	1.68	3.61
Ever-use of contraception	1.05	1.12	1.48
No correlated variance	2.00	3.33	10.00

As we would expect, the variables *Children ever born* and *Age at marriage* have values of V_c/V_t close to the upper limit. This is because there is no interviewer variance for these variables and the correlated sampling variance is relatively small. The results for the variable *First birth interval* show how unwise it would be to apply this limit to a case where either of the correlated variance components is reasonably large. Under the assumption of this model, using the upper limit for the variance would lead to over-estimating the total variance by about half when $M_c = 0.5$; by about 100 per cent when $M_c = 0.3$; and by nearly 300 per cent (177 per cent for Peru) when $M_c = 0.1$.

The last variable in the table shows even more dramatic results. This variable is atypical since both the correlated sampling variance and the interviewer variance are extremely large. In such a situation, however, the effects are astonishing. For a crossclass with $M_c = 0.5$ the

total variance is almost identical to the total variance for the whole sample, although the sample size for the subclass is only half the size of the whole sample. The further reduction of sample size for M_c = 0.3 and M_c = 0.1 leads only to relatively small increases in the variance. For M_c = 0.1 (a crossclass comprising one tenth of the sample) the ratio of V_c/V_t is less than 2 (1.82 for Lesotho and 1.48 for Peru). For a variable with no correlated variance this ratio would be 10.00.

The results in the previous tables can also be presented in a form closer to the approach used in discussing sampling variance. Tables 5.12L and 5.12P give the values of *deff, inteff* and *toteff*, where

$$toteff = deff + inteff - 1 \tag{5.13}$$

and *toteff* is the ratio of the *total variance* to the *simple total variance.*

Table 5.12L Values of *deff, inteff* and *toteff* for different values of M_c for Lesotho

Variable	Measure	Total sample	M_c = 0.5	M_c = 0.3	M_c = 0.1
Age at marriage	deff	1.00	1.00	1.00	1.00
	inteff	1.00	1.00	1.00	1.00
	toteff	1.00	1.00	1.00	1.00
Children ever born	deff	1.14	1.07	1.04	1.01
	inteff	1.00	1.00	1.00	1.00
	toteff	1.14	1.07	1.04	1.01
First birth interval	deff	1.44	1.21	1.12	1.04
	inteff	4.46	3.21	2.31	1.41
	toteff	4.90	3.42	2.43	1.45
Ever-use of contraception	deff	2.37	1.66	1.38	1.14
	inteff	8.32	5.12	3.44	1.76
	toteff	9.69	5.78	3.82	1.90

Table 5.12P Values of *deff, inteff* and *toteff* for difference values of M_c

Variable	Measure	Total sample	$M_c = 0.5$	$M_c = 0.3$	$M_c = 0.1$
Children ever born	deff	1.14	1.06	1.04	1.00
	inteff	1.00	1.00	1.00	1.00
	toteff	1.14	1.06	10.4	1.00
Age at marriage	deff	1.10	1.04	1.02	1.00
	inteff	1.00	1.00	1.00	1.00
	toteff	1.10	1.04	1.02	1.00
First birth interval	deff	1.30	1.16	1.08	1.01
	inteff	2.99	1.99	1.59	1.18
	toteff	3.29	2.12	1.66	1.19
Ever-use of contraception	deff	3.39	2.10	1.59	1.08
	inteff	11.02	6.00	4.00	1.91
	toteff	13.41	7.03	4.51	1.98

The results in tables 5.12L and 5.12P conform to the pattern observed in the sampling literature for crossclasses. Under the assumptions of the model the effect of the correlated variance components decreases as the proportion of the population in the crossclass decreases. The larger the effect of the correlated variance components, the more dramatic the reduction as M_c decreases. The rates at which *deff* and *inteff* decrease differ as the multipliers *(b-1)* and *(m-1)* differ.

Finally, to illustrate the practical implications of these results for the evaluation of survey estimates, tables 5.13L and 5.13P give the width of the 95 per cent confidence intervals for cross-classes of different sizes. The same four variables are presented and the width of the confidence interval for the estimate based on the whole sample is also given for comparison.

Table 5.13L Width of 95 per cent confidence intervals for cross-classes of different sizes for Lesotho

Variable	Mean	Cross-class	Simple sampling error (1)	Total simple error (2)	(2)×deft (3)	Correct standard error (4)
Age at marriage	17.90	Total sample	0.173	0.208	0.208	0.208
		$M_s = 0.5$	0.245	0.294	0.294	0.294
		$M_s = 0.3$	0.316	0.380	0.380	0.380
		$M_s = 0.1$	0.547	0.658	0.658	0.658
Children ever born	3.19	Total sample	0.162	0.169	0.180	0.180
		$M_s = 0.5$	0.229	0.239	0.246	0.246
		$M_s = 0.3$	0.296	0.309	0.314	0.314
		$M_s = 0.1$	0.512	0.534	0.535	0.535
First birth interval	25.96	Total sample	1.00	1.55	1.86	3.43
		$M_s = 0.5$	1.414	2.192	2.411	3.74
		$M_s = 0.3$	1.826	2.830	3.000	4.13
		$M_s = 0.1$	3.162	4.901	4.999	5.66
Ever-use of contraception	0.23	Total sample	0.0172	0.0256	0.0394	0.0797
		$M_s = 0.5$	0.0243	0.0362	0.0467	0.0832
		$M_s = 0.3$	0.0314	0.0467	0.0546	0.0877
		$M_s = 0.1$	0.0544	0.0809	0.0866	0.1075

The relationship between the standard error for a subclass and the standard error for the whole sample is determined by two factors: (i) the size of the sample for the subclass; the smaller the sample size (ie the smaller M_c) the larger the standard error will be - this applies to all components of the total variance; (ii) the relative size of the correlated errors; in the absence of correlated errors, the only influence will be the relative sizes of the total sample and the subclass. However when there are correlated errors, either sampling or response, the relationship becomes more complex. For crossclasses, the model described implies that there will be a considerable dilution of the effect of the reduction in sample size. This is because the impact of the correlated errors depends critically on the size of the `clusters' within which the errors are correlated; for the correlated sampling errors the cluster take is the dominant factor, for correlated interviewer errors the interviewer workload size is the critical consideration. For small crossclasses both these sizes are greatly reduced, with a consequent

reduction in the correlated components.

Table 5.13P Width of 95 per cent confidence intervals for crossclasses of different sizes for Peru

Variable	Mean	Subclass	Simple sampling error (1)	Total simple error (2)	(2)×deft	Correct standard error (4)
Children ever born	4.66	Total sample	0.356	0.360	0.354	0.384
		$M_s = 0.5$	0.504	0.508	0.524	0.524
		$M_s = 0.3$	0.648	0.656	0.668	0.668
		$M_s = 0.1$	1.124	1.136	1.136	1.136
Age at marriage	19.9	Total sample	0.428	0.476	0.500	0.500
		$M_s = 0.5$	0.604	0.672	0.684	0.684
		$M_s = 0.3$	0.776	0.872	0.880	0.880
		$M_s = 0.1$	1.348	1.508	1.508	1.508
Firth birth interval	11.3	Total sample	1.40	2.12	2.40	3.85
		$M_s = 0.5$	1.968	2.972	3.212	4.389
		$M_s = 0.3$	2.540	3.836	3.992	5.005
		$M_s = 0.1$	4.400	6.644	6.676	7.315
Ever-use of contraception	0.56	Total sample	0.0460	0.0572	0.1052	0.2094
		$M_s = 0.5$	0.0652	0.0812	0.1176	0.2136
		$M_s = 0.3$	0.0840	1.1048	0.1320	0.2220
		$M_s = 0.1$	0.1456	0.1812	0.1884	0.2555

The final columns of tables 5.13L and 5.13P encapsulate the results of this section. For the two variables *Children ever born* and *Age at marriage* the ratio of the standard errors (and thus of the confidence intervals) is close to that expected on the basis of sample size alone - the correlated errors are relatively unimportant. For the *First birth interval* the confidence interval for the smaller crossclasses is a good deal narrower than would be expected if sample size were the only consideration. For *Ever-use of contraception* the confidence interval for the crossclass with $M_c = 0.1$ (ie based on one tenth of the total sample) is only 22 per cent wider than the confidence interval based on the total sample. This is because the dominance of the correlated error in the standard error of estimates based on the total sample becomes progressively weaker as the crossclass size decreases as a proportion of the total sample size. The standard error for the crossclass is still larger than it would be if the correlated variance

components were all zero, but the relative increase in the standard error is much smaller than for the whole sample.

5.3.3 Discussion

A model is presented which describes the total variance of an estimate in terms of five factors: the simple total variance; the synthetic intracluster correlation coefficient for the sample design; the synthetic intracluster correlation coefficient for the fieldwork design; the average cluster take; and the average interviewer workload size. The model is analogous to that generally used to describe the total sampling variance. The implications of this model for the total variance of estimates based on cross-classes were presented and a simple expression was derived for the relationship between the total variance for the total sample and the total variance for a cross-class. A number of important assumptions are made in the model. First, it is assumed that the cross-classes are uniformly distributed across clusters and interviewers; in the context of WFS surveys, age subclasses are likely to satisfy this condition at least approximately. Secondly, it is assumed that the intra-cluster and intra-interviewer correlation coefficients remain constant for cross-classes. The evidence on this is less convincing, although it seems a useful approximation in practice. In particular, the evidence for the intra-cluster correlation coefficient suggests that it is reasonably stable. Further investigation of the behaviour of the intra-interviewer correlation is desirable.

The applications illustrate the implications of the theoretical model. The results are presented for Lesotho and Peru for four variables which represent the different situations which might arise; the pattern of the results is strikingly similar for the two countries. For two of the variables the total variance is primarily due to the simple sampling variance and the simple response variance. In this case the relative size of cross-class variance is determined largely by the cross-class size. For the third variable there is a more substantial correlated sampling variance component and also a correlated response variance component. The total correlated variance dominates the total variance for estimates based on the whole sample. However, for cross-classes this dominance is reduced as the cross-class size decreases. For small crossclasses the simple variance predominates and the effects of the correlated variances almost disappear. The situation is even more striking for the fourth variable - the total effect

(the ratio of the total variance to the simple total variance) is 9.7 for Lesotho (13.4 for Peru) for estimates based on the total sample and only 1.90 (for Lesotho, 2.0 for Peru) for estimates based on a cross-class representing one-tenth of the total sample. This is an extension of the results obtained for sampling variance in other studies - the effects of the design are diminished as the size of the cross-class is reduced.

Although the results are not based on direct computations of the variance, values of the parameters on which the calculations are based are obtained from computations carried out on data from the Lesotho and Peru studies. There is a problem in estimating the correlated variance components for the cross-classes in that as the sample size decreases the estimates themselves become subject to larger variances.

Where the correlated interviewer variance is large this problem is less severe. For Lesotho the variable with the largest correlated interviewer variance was *Ever-use of contraception*. Table 5.14L gives the estimated values of ρ_i for five subclasses for this variable, together with their estimated standard errors. The computations are for the data from the main survey.

Table 5.14L Estimated values of ρ_i for cross-classes for *Ever-use of contraception* for Lesotho

	Total sample	Age < 25	Age 25-34	Age 35-44	Educ. 1-5yr	Educ. 6yr +
ρ_i	0.084	0.061	0.121	0.072	0.088	0.052
$se(\rho_i)$	0.018	0.042	0.038	0.044	0.034	0.024
M_c	1.00	0.31	0.33	0.27	0.44	0.48

From the table it can be seen that the values of ρ_i for the subclasses are consistent with the assumption that ρ_i remains constant across subclasses. In no case is the value of ρ_i more than one standard error from the value of 0.084 obtained for the total sample. This evidence provides some support for the model.

The choice of sample design and field design for a survey tends to be determined by material

and practical constraints imposed by the data collection operation. Nevertheless, data relating to sampling and response errors can provide a more rational basis for making decisions about the design. The findings of this section, however, illustrate a particular difficulty. A basic consideration in evaluating the design is the relative importance attached to estimates based on the whole sample compared with those for sample subclasses and subclass differences. Generally, the smaller a subclass the less sensitive is the associated variance to specific features of the design. In particular, the less is the effect of the correlated components of the variance and the more ill-defined is the `optimal' solution to the problem of survey design. A weighted loss function would be necessary to incorporate these considerations in design optimization.

Chapter 6 - THE UNDERLYING RELIABILITY (QUALITY) OF THE DATA FROM THE WFS

The purest measure of reliability is the difference between the values obtained for repeated observations on the same individuals (ideally at the same time). In the practical survey situation the repeated observations must necessarily be separated in time. The data can however be edited to apply to the same time point (or time period), though there is of course an implicit assumption that the two observations are independent of each other. In general it is impossible to test this assumption; two approaches to the problem appropriate for particular cases are presented in chapters 8 and 9.

In both the Peru and Lesotho surveys a subsample of respondents was reinterviewed some time after the main survey. The responses from the reinterviews were coded to refer the same time period etc as the original interview. These two observations therefore constitute a set of repeated measurements that we will treat as independent. Section 6.1 considers the simplest approach to the measurement of reliability - crosstabulation of the responses from the two occasions. Section 6.2 examines some elementary measures of agreement between the two sets of observations. Section 6.3 uses the simple response variance (SRV) as a measure of reliability. Section 6.4 looks at the extent to which the SRV varies across subclasses of the sample.

6.1 Crosstabulation of Repeated Observations

For each individual interviewed in the re-interview survey we have two separate observations for each variable. The differences within and between the pairs of observations provide the raw material for the investigation. In general, reliability can be defined as the extent to which a measurement remains constant as it is repeated under conditions taken to be constant. Thus a useful measure of reliability should take into account variations in the individual observations. At a basic level, the most illuminating presentation is that which describes the set of deviations between the observations on the two occasions. This approach has the further advantage that it applies to all types of variables and that the magnitudes of the individual response deviations can be interpreted

substantively. In addition, it is applicable to the whole set of variables, regardless of the level of measurement - nominal, ordinal or metric.

In this section we consider some examples of this basic procedure. In examining the responses obtained on the two occasions for a particular variable, the data can be represented by a cross-classification of the two sets of responses. Tables 6.1-6.4 are examples of such cross-tabulations. Corresponding tables are presented for Peru and Lesotho.

Table 6.1P presents the data for the variable *Ever-use of contraception* for Peru; table 6.1L the corresponding table for Lesotho. This is a binary variable and thus all the information is contained in a simple 2×2 table.

Table 6.1P Ever-use of contraception as reported in the original interview and the re-interview in Peru

Original interview	Re-interview		
	Yes	No	Total
Yes	591	184	704
No	41	453	494
Total	560	638	1198

For both Peru and Lesotho approximately 20 per cent of the women gave inconsistent responses on the two occasions. The observed response variability stems at least in part from the fact that the basic condition of comparability - the `essential survey conditions' being the same for the two interviews - was violated. The method of questioning in the two interviews differed. In the original interview, the respondent was asked to name the contraceptive methods she had `heard of' and for each method mentioned she was asked whether she had ever used it; this was followed by the interviewer's reading out a description of a number of other methods one by one and repeating the question on use in

each case. This extra probing was not done in the second interview, and a substantial proportion of respondents may consequently have failed to report contraceptive use. The level of ever-use of contraception reported in the first interview was 12 per cent higher than in the second interview (14 per cent in Lesotho), with 15 per cent of all respondents reporting use in the original interview and not in the re-interview (16 per cent in Lesotho), whereas only 3 per cent reported use in the re-interview and not in the original interview.

Table 6.1L Ever-use of contraception as reported in the original interview and the re-interview in Lesotho

Original interview	Re-interview		
	Yes	No	Total
Yes	67	99	166 (27.3%)
No	16	427	443 (72.7%)
Total	83 (13.7%)	526 (86.3%)	609

These tables illustrate two strengths of this direct presentation. First, a comparison of the marginals provides an indication of whether there is any major difference between the results of the two interviews for the whole sample, which serves as a check on the constancy of the essential survey conditions. Secondly, the cells of the cross-tabulations give a vivid picture of the scale of the response deviations for the individual respondents.

Table 6.2P presents the cross-tabulation for *Level of education* for Peru. This is an ordinal variable where increasing values of the categories represent greater exposure to education.

Table 6.2P Educational level as reported in the original interview and the re-interview in Peru

Original	Re-interview					
interview	1	2	3	4	5	Total
1	259	46	0	4	0	309
2	27	191	18	1	1	238
3	4	35	115	9	1	164
4	1	3	24	178	3	209
5	0	4	2	12	260	278
Total	291	279	159	204	265	1198

This table provides a greater wealth of detail than table 6.1 because of the number of categories involved. The categories are also in rank order and the difference between the categories is of substantive significance. By observing the marginals of the table, we see that the pattern of results is broadly similar for the two interviews.

The level of education reported differed for one in six respondents, ie for 195 women. For the great majority of those - 174 - the difference between the two responses amounted to a shift through one educational level; for 11 women the discrepancy was two educational levels; and for the remaining 10 women there was a difference of three levels.

Tables 6.2LA and 6.2LB present the cross-classification for *Education* in Lesotho. In the case of table 6.2LA the data are presented for the categories of educational level. The categories are in rank order and the differences between them are of substantive significance.

Table 6.2LA Educational level as reported in the original interview and the re-interview in Lesotho

Original interview	Re-interview			
	1	2	3	Total
1	157	31	3	191
2	29	192	32	253
3	4	42	119	165
Total	190	265	154	609

The level of education reported differed for almost one in four respondents, ie for 141 women. For the great majority of those - 134 - the difference between the two responses amounted to a shift through one educational level. In only seven cases was there a shift through two levels. By observing the marginals of the table, we see that the pattern of results is broadly similar for the two interviews, providing some reassurance that for this variable the essential survey conditions remained constant.

In table 6.2LB the data are presented for number of years' education completed. This is a metric variable for which the size of the discrepancy in each case has a clear and unambiguous meaning. For 63 per cent of the respondents the two observations agree. For only 11 per cent did the discrepancy exceed one year. The pattern of the marginals is still broadly similar - the original interviews and the re-interviews produce comparable distributions of years of education completed. Table 6.2LB provides more information about the two sets of responses than does table 6.2LA but the additional detail also has the effect of making the information more difficult to assimilate.

Table 6.2LB Education in years as reported in the original interview and the re-interviews in Lesotho

Original interview	Re-interview											
	0	1	2	3	4	5	6	7	8	9	10	Total
0	41	2	0	1	1	1	0	0	0	0	0	46
1	0	2	1	2	0	1	0	0	0	0	0	5
2	2	2	4	2	0	1	1	1	0	0	0	12
3	2	0	8	23	6	7	3	2	0	0	0	50
4	2	1	3	7	46	9	9	0	0	0	0	77
5	0	0	1	1	22	76	15	5	0	0	0	120
6	1	0	1	2	1	16	85	27	0	0	0	133
7	0	1	0	1	0	8	32	93	2	0	0	137
8	0	0	0	0	0	0	2	3	2	2	0	9
9	0	0	0	0	0	0	0	1	4	6	0	11
10	0	1	0	0	1	0	0	0	0	1	5	8
Total	47	9	18	39	77	118	147	132	8	9	5	609

Tables 6.3P and 6.3L deal with one of the variables of central importance in a fertility survey - the *Number of children ever born* to the respondent (parity). The data from Peru show better quality overall than Lesotho.

For both tables, partly because of the size of the table (the number of categories), the pattern of results is striking. For the great majority of respondents the responses on the two occasions are identical. For a variable that seems as unequivocal as this, however, it is perhaps surprising that any observations differ on the two occasions. Most of the discrepant cases involve a difference of only one, but much larger deviations are found in a number of cases (as large as four in Peru, and as large as seven in Lesotho); these cases may involve reporting/non-reporting of children from other unions.

Overall one in eight women reported inconsistently in Peru, one in five in Lesotho; the reporting is less consistent at higher parities than at lower parities, as might be expected. Nevertheless the marginal distributions are very similar, and the means for the original

interview and the re-interview are almost identical.

The problems of providing a useful summary of the data are illustrated by these tables. For Peru, there are 361 cells in the table, of which 342 would indicate a discrepancy between the two determinations; only 46 of these cells contain observations. For Lesotho, there are 196 cells in the body of the table, of which 182 would indicate a discrepancy between the determinations; only 54 of these cells contain observations. The importance of each of these cells depends on the size of the discrepancy it represents and the number of cases in the cell. To discuss each of the occupied cells in turn, however, would be both lengthy and uninformative. This problem is exacerbated by the fact that we wish also to describe the reliability of the data for subclasses of the sample. Thus we will certainly be forced to condense the tables into some summary measures which contain the information necessary to evaluate the data.

Table 6.3P Number of children ever born as reported in the original interview and the re-interview for Peru

Original interview	Reinterview																				
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Total
0	40	5																			46
1	2	124	7	3																	136
2		1	148	8																	158
3			3	151	11	4	1														169
4			1	4	139	7	1														151
5					2	107	8	4		1											121
6					4	4	98	10													116
7						6	3	78	3												89
8							1	5	55	7	1		1								69
9								1	1	38	5										45
10									4		21	2	1								28
11								3		1	5	26	3								38
12												1	11								11
13												1	1	3	1						5
14											3			1	5						9
15												1				3					4
16																	2				2
17																					
18																					
19																				1	1
Total	43	131	159	165	156	177	111	100	62	47	34	30	17	4	6	3	2	0	0	1	1198

Table 6.3L Number of children ever born as reported in the original interview and the re-interview for Lesotho

Original interview	Re-interview														
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	Total
0	64	5		1			1	1							72
1	2	92	12	4	1	1	1		1						114
2	2	4	84	4	2		1	1	1						99
3	1	1	2	67	5	1	1								78
4	1	1		1	52	11			1						67
5			1		2	46	6			1					56
6				1	1	5	28	6	1			1			43
7	1					1	3	24	3	1					33
8		1		2				3	11	1					18
9									3	11	1				15
10									1	1	5	2	2		11
11												1			1
12													2		2
13														1	1
Total	71	104	99	80	63	65	41	35	22	15	6	4	4	1	609

One further variable may be considered here to illustrate the difficulty. Tables 6.4P and 6.4L give the two sets of responses for one of the few attitudinal variables included in most WFS national surveys - *Number of children desired.* This table is dramatically different from tables 6.3P and 6.3L. We would expect an attitude variable to be particularly subject to response variability and the tables confirm this expectation. Furthermore, this variable is different in kind from the variables considered in tables 6.1 to 6.3 in that the true value of the variable may change between the two interviews.

In fact, fewer than half the women gave identical responses on the two occasions (only two in five in Lesotho). The discrepancies are large and the overall impression is of very unreliable reporting. From a substantive point of view, this variable is of interest more as an indication of the desire for small or large families rather than as a precise measure of behaviour, and it is encouraging that 70 per cent (Peru) and 60 per cent (Lesotho) report the number desired

within one child in the two interviews. The marginal distributions are relatively stable and the means of the two distributions are very close.

The tables presented in this section illustrate both the strengths and the weaknesses of this kind of analysis. For the subject matter specialist it is clearly important to look in detail at the pattern of individual response deviations. The only satisfactory way of doing this is to cross-tabulate the two sets of responses for each variable. Tables such as tables 6.1 through 6.4 provide an opportunity to examine the deviations in the context of the values obtained from the two interviews and thus allow the analyst to investigate the underlying response process. But the tables are relatively unwieldy and cannot in practice be presented and examined for every variable for every subclass of interest. It is therefore necessary to consider how the information may be condensed and summarized to make it more manageable and more easily interpretable. There is conflict between detail and assimilation. In the next section the simplest summary measures are presented.

Table 6.4P Number of children desired as reported in the original interview and the re-interview for Peru

Original interview	Reinterview															
	0	1	2	3	4	5	6	7	8	9	10	11	12	98	99	Total
0	1	2	1	2	3	1	1								1	11
1	1	10	19	1	4	0	0						1		9	45
2	3	12	134	46	35	7	4	1			1	1	0	2	9	253
3	6	3	62	129	48	14	13	0	5		0	0	1	5	12	298
4	1	6	31	34	145	10	25	1	1		1	0	2	1	10	268
5		2	15	18	11	16	9	1	5		3	0	0	6	10	96
6		1	7	21	16	16	29	2	3	1	1	1	4		9	110
7			0	1	1	3	2	2	3	0	0		0		3	15
8			1	4	3	7	4	1	2	0	0		0		3	25
9			0			0	2			0	0		0		6	8
10	1		3			1				1	3		1		5	15
11																0
12				2		1	0		1		4		7		1	10
16						1										1
20			3													3
98	1		3	4	1					1					6	15
99			3	7	1	4								3	8	25
Total	12	36	281	270	268	81	88	7	19	1	12	2	10	17	93	1198

Table 6.4L Number of children desired as reported in the original interview and the re-interview for Lesotho

Original Interview	Re-interview																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	16	17	98	Total
1	1		1			1				1								4
2		5		1	1	1	2	1										11
3		4	12	10	2	8		2		1	1							40
4		4	6	68	18	18	5	4		5	1						4	136
5		2	2	6	28	13	3	2	3	5		1			1		6	70
6	1	2	3	22	7	60	11	9	1	7	2	4	1	1			4	134
7	1			3	1	7	16	6	1	4	1	1		1			7	50
8			2	2	2	11	3	22	2	11		1					1	59
9				1		1		1	5	2								11
10		1	1	3	2	4	1	5	6	33	1	6	3			1	4	66
11						1			1		0							1
12										3	1	3		1				9
13									1			1	1					2
14			1				1							1				3
20				1													1	2
98			1	1	1	1		3		1							4	12
Total	3	18	29	118	62	125	41	55	20	73	6	18	5	4	1	1	30	609

6.2 Simple Summary Measures of Quality

The measures used in this section are defined in section 2.3.2 of chapter 2. Tables 6.5P and 6.5L present the values of D, A, κ and κ_w for eighteen variables. For most variables the index of crude agreement, A, is very close to the supposedly more refined measure κ. This is probably due to the fact that for most of the variables considered the number of categories involved is large, with no dominant category. For an approximately uniform distribution across a large number L of categories, $p_e = 0(1/L)$ and it follows from equations (2.27) and (2.28) that for a reasonably consistent set of data, $A = p_o >> p_e$ so that $\kappa \doteq A$. Hence little is gained by introducing κ in such cases.

Table 6.5P Values of D, A, κ and κ_w (values x 100) for Peru

Variable	D	A	κ	κ_w
Level of education	16	84	79	94
Children ever born	12	88	86	98
Ever-use of contraception	19	81	63	63
Age	46	54	53	98
Age in 5 year groups	15	85	83	97
Age at marriage	54	46	41	79
Year of marriage	49	51	50	96
Marital duration (years)	57	43	41	96
Births in past 5 years	16	84	78	91
No. of children desired	56	44	31	42
First birth interval (months)	65	35	32	43
Last closed birth interval (months)	55	45	44	76
Year of first birth	29	71	70	97
Year of last birth	25	75	72	97
Year of next to last birth	36	64	61	94
Status of first union	05	95	83	-
Worked since marriage (binary)	13	87	70	70
Worked since marriage	28	72	66	-

Table 6.5L Values of D, A, κ and κ_w (values × 100) for Lesotho

Variable	D	A	κ_w	κ
Years of education	35	65	58	87
Children ever born	19	81	78	92
Ever-use of contraception	19	81	42	42
Current age	40	60	58	94
Age in five-year groups	17	83	80	93
Age at marriage	51	49	43	69
Year at marriage	29	71	70	93
Marital duration (years)	40	60	59	93
Births in past five years	18	82	74	85
No of children desired	54	46	36	51
First birth interval (months)	52	48	47	41
Last closed birth interval (months)	50	50	49	66
Year of first birth	25	75	74	94
Month of first birth	36	64	64	94
Year of last birth	23	77	73	96
Month of last birth	35	65	65	96
Year of next to last birth	28	72	69	94
Month of next to last birth	44	56	56	94

In terms of the measures used in the tables 6.5P and 6.5L the four most unreliable variables are the *Number of children desired, Age at marriage*, the *First birth interval* and the *Last closed birth interval.* The last three of these are composite variables derived from two or more questions, each of which is subject to error. The other is one of the few attitudinal items in the questionnaire, and may be expected to be particularly sensitive to response variability (the full cross-tabulation of the responses for this variable is given in tables 6.4P and 6.4L). Even so, the degree of unreliability gives cause for concern. Fully half of the respondents gave different responses on the two occasions, and the correlation between the two sets of responses is between 0.4 and 0.7.

For a further set of variables the level of disagreement between the responses is also high, with about one-third of the individuals giving inconsistent responses. These are variables essentially dependent on single dates reported in the interviews.

Among the variables least affected by response variability are the two measures of fertility which are central to much of the WFS analysis. These are the number of *Children ever born* and *Births in past 5 years*. This is reassuring, although even for these variables the responses from the two interviews are by no means perfectly consistent. Another variable which performs well is *Age in 5-year groups*. The difference between the apparent reliability of *Age* and *Age group* arises from the fact that many of the discrepancies in the age variable nevertheless do not cause the individual to cross the boundary of the age group. It is worth noting that even for this variable one in six of the women is classified in a different age group in the two interviews.

One further point may be worth noting. The dichotomies included in the tables (two in table 6.5P and one in table 6.5L) - *Ever-use of contraception* and *Worked since marriage* - perform reasonably well except in terms of κ_w. Since these variables have only two categories, each discrepancy receives considerable weight in the computation of κ_w. This is appropriate since a discrepancy represents a complete misclassification on one occasion. These cases may be contrasted with the converse cases where the value of κ_w is high relative to A and κ; *Age, Children ever born, Age at marriage* and *Last closed birth interval* are examples. This is because in these cases the discrepancies, though they may be substantively serious, are small in relation to the possible range of values for the variable.

The measure κ_w does not seem from the table to be particularly useful. Where the range of the variable is very wide, as it is for many of these variables, the discrepancies, while substantively serious, are small in comparison. In such cases κ_w is a rather insensitive index of consistency. Furthermore, since the marginal distributions are in general fairly close, κ_w will tend to be almost identical to the correlation between the two sets of responses (ie those for the original interviews and the re-interviews).

Part of the difficulty in evaluating tables 6.5P and 6.5L arises from the fact that the measures considered in this section do not fit easily into the framework of survey analysis and are either too crude, as in the cases of A and D, or unsatisfactory in terms of substantive interpretation,

as with κ and κ_w. In section 6.3 a more general approach is described.

6.3 The Components of the Simple Total Variance

The simple variance estimated from a sample of observations automatically includes the simple sampling variance and the simple response variance. With repeated observations we obtain in effect two estimates of this simple total variance, one from the original interviews and one from the re-interviews. The simple sampling variance and the simple response variance, however, can only be estimated from the two sets of observations together. This section gives examples of the estimation of the components of the simple total variance and of the index of consistency, *I*. Similar analyses are presented first for both Peru and Lesotho. A more extensive analysis of the data from Lesotho follows. Finally the results for Lesotho and Peru are summarized in figures 6.1L and 6.1P.

Tables 6.6P and 6.6L present the data for three frequency distributions: the responses in the original interview, the responses in the re-interviews, and the (case-by-case) deviations between the two sets of responses. On the basis of these data, the parameters of the three frequency distributions can be estimated. The distributions of the responses in the original interviews and the re-interviews provide estimates of the simple total variance. The distribution of the deviations provides an estimate of the simple response variance. Tables 6.7P and 6.7L give the estimates.

6.3.1 An example for Peru

Table 6.6P Data for the estimation of simple variance components for the variable *Births in the past five years* for Peru

Value	Original interview	Re-interview	Deviation between original interview and re-interview	Frequency
0	407	440		
1	389	381	-2	4
2	283	266	-1	58
3	103	97	0	1009
4	16	10	1	120
5	0	4	2	7
Total	1198	1198	Total	1198

Table 6.7P Components of the simple variance for *Births in the past five years*

	Original interviews	Re-interviews	Deviations
Mean	1.110	1.054	0.056
Standard deviation	1.018	1.017	0.427
Standard error	0.029	0.029	0.012
Variance	1.037	1.034	0.182

Both 1.037 and 1.034 are estimates of the simple total variance $\sigma_y^2 + \sigma_\varepsilon^2$, whereas 0.182 is an estimate of $2\sigma_\varepsilon^2$. Thus the estimate of σ_ε^2 is

$$\hat{\sigma}_{\varepsilon}^2 = \tfrac{1}{2}(0.182) = 0.091$$

The best available estimate of $\sigma_y^2 + \sigma_{\varepsilon}^2$ is

$$\hat{\sigma}_y^2 + \hat{\sigma}_{\varepsilon}^2 = \tfrac{1}{2}(1.037 + 1.034) = 1.0355$$

Consequently

$$\hat{I} = \frac{0.091}{1.0355} = 0.08788$$

This procedure makes use of all the available data. Instead of using the matrix containing the full cross-classification of the responses from the two interviews (examples are given in tables 1-4 in section 4) which becomes unwieldy when the number of categories is large, the data are used in the form given in table 6.6P. All the components of the simple total variance can be derived from these distributions.

6.3.2 Two examples for Lesotho

Table 6.6L Data for estimation of simple variance components for two variables

Value	Original interview	Re-interview	Deviation between original interview and re-interview	Frequency
A *Births in the last five years*				
0	215	204	-3	1
1	217	209	-3	4
2	155	168	-2	69
3	22	26	0	501
4	1	3	1	30
			2	4
			3	1

B *Ever-use of contraception*				
0	446	531	-1	13
1	163	78	0	498
			1	98

Table 6.7L Components of the simple variance for two variables

Measure	Original Interviews	Re-interviews	Deviations
A *Births in the last five years*			
Mean	0.977	1.041	-0.063
Standard deviation	0.884	0.914	0.493
Standard error	0.036	0.037	0.020
Variance	0.782	0.836	0.243
B *Ever-use of contraception*			
Mean	0.267	0.129	0.138
Standard deviation	0.445	0.333	0.412
Standard error	0.018	0.014	0.017
Variance	0.198	0.111	0.170

From table 6.7L(A), both 0.782 and 0.836 are estimates of the simple total variance, whereas 0.243 is an estimate of the simple response variance multiplied by two. Thus the estimate of σ_ε^2 is

$$\hat{\sigma}_\varepsilon^2 = \tfrac{1}{2}(0.243) = 0.1215$$

The best available estimate of $\sigma_y^2 + \sigma_\varepsilon^2$ is

$$\sigma_y^2 + \sigma_\varepsilon^2 = \tfrac{1}{2}(0.782 + 0.836) = 0.809$$

In order to obtain a single estimation of σ_y^2 from the two sets of observations σ_y^2 is obtained by subtracting $\hat{\sigma}_\varepsilon^2$ from the estimate of $\sigma_y^2 + \sigma_\varepsilon^2$.

Thus,

$$\hat{\sigma}_y^2 = \tfrac{1}{2}(0.782 + 0.836) - \tfrac{1}{2}(0.243) = 0.6875$$

From table 6.7L(B), the corresponding estimates for ever-use of contraception are:

$$\hat{\sigma}_\varepsilon^2 = \tfrac{1}{2}(0.170) = 0.085$$
$$\sigma_y^2 + \sigma_\varepsilon^2 = \tfrac{1}{2}(0.198 + 0.111) = 0.1545$$
$$\hat{\sigma}_y^2 = 0.1545 - 0.085 = 0.0695$$

The estimates above provide all the information required for the estimation of the index of inconsistency, *I*. For births in the past five years:

$$\hat{I} = \frac{0.1215}{0.809} = 0.150.$$

For ever-use of contraception:

$$\hat{I} = \frac{0.085}{0.1545} = 0.550$$

6.3.3 Further analysis for Lesotho

Table 6.8L presents the components of the simple total variance for eighteen key variables for Lesotho, again arranged in order of increasing values of *I*. The variables show a very wide range of values, but some interesting conclusions may be drawn from the table.

Table 6.8L Components of the simple response variance for eighteen variables for Lesotho

No	Variable	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	$\hat{I}$
1	Year of last birth	29.0600	1.1180	0.037
2	Month of last birth	4210.0000	171.3000	0.039
3	Year of next to last birth	28.7700	1.6490	0.054

4	Month of next to last birth	4147.0000	244.0000	0.056
5	Year of first birth	67.7900	4.4560	0.062
6	Month of first birth	9752.0000	631.7000	0.061
7	Age	78.5500	5.1680	0.062
8	Age in five-year groups	3.1220	0.2325	0.070
9	Year of marriage	74.8800	5.9730	0.074
10	Marital duration	75.1800	5.9640	0.074
11	Children ever born	6.1800	0.5580	0.083
12	Years of education	4.2770	0.6460	0.131
13	Births in past five years	0.6875	0.1215	0.150
14	Age at first marriage	7.0170	3.1020	0.306
15	Last closed birth interval	324.9000	185.8000	0.364
16	No of children desired	3.5230	3.2480	0.480
17	Ever-use of contraception	0.0695	0.0850	0.550
18	First birth interval	234.9000	330.8000	0.585

In assessing the meaning of the values of $\hat{I}$ presented in table 6.8L it is important to bear in mind that $\hat{I}$ is a ratio of two variances. The numerator $\hat{\sigma}_{\varepsilon}^2$ is the simple response variance and is a measure of the magnitude of the inconsistencies in the responses. The denominator $\hat{\sigma}_{y}^2 + \hat{\sigma}_{\varepsilon}^2$ is the simple total variance, which measures the total variability in the observations. The size of the ratio $\hat{I}$ therefore depends critically on the size of $\hat{\sigma}_{y}^2$, and values of $\hat{I}$ may be misinterpreted unless the analyst is aware that the value of $\hat{\sigma}_{y}^2$ may well be different even for variables which are measured in the same units (*age* and *age at marriage*, for instance). The index $\hat{I}$ measures the proportion of the total variability in the responses which is due to disturbances introduced into the observations by the measurement process itself.

Somewhat surprisingly, the six variables with the lowest values of I are the individual dates obtained for the first, next to last and last births. The values of I range from 0.037 to 0.062. There is a consistent increase in I as the births become more distant in time. In fact, the deterioration in reporting is more severe than that indicated by I itself. An inspection of the simple response variance $\left(\hat{\sigma}_{\varepsilon}^2\right)$ shows that for the dates in years the value rises from 1.1 through 1.6 to 4.5 and for the dates in months from 171 through 244 to 632. The relatively slight increase in I for the first birth is due to the fact that the possible range of values (and the corresponding variance) for *Date of first birth* is considerably wider for the total sample of women aged 15-49 than the range of values for *Date of last birth*. However, overall, these six variables seem to be reliably reported.

The next four variables are central to much of the analysis of WFS data. Age, age in five-year groups, year of first marriage and marital duration are all used widely as classification variables. All four have values of I near 0.07 and seem on this basis to be measured reliably. This issue, however, is considered in more detail in the following sub-section.

The three variables following the date-based variables have moderate values of I. Two of them - *Children ever born* and *Births in the past five years* - are variables of crucial importance to the analysis of fertility. It may seem a little surprising that the more recent data have a larger value of I than the more general variable *Children ever born*. This arises from two factors. First, births in the past five years depends not only on births being reported but also on the dates of these births, thus adding a potential source of inconsistency. Secondly, the value of $\hat{\sigma}_y^2$ is much larger for children ever born, so that although the simple response variance is much lower for births in the past five years, the value of I is almost twice as large.

The remaining five variables have high values for I. In each case the proportion of the simple total variance due to the simple response variance is over 30 per cent. Three of the variables are based on dates and it is remarkable that the values of I are so large given the apparent

reliability of the individual date variables. Table 6.9L gives the relevant data.

Table 6.9L Results relating to the reliability of age at first marriage, last closed birth interval and first birth interval for Lesotho

Variable	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	$\hat{I}$	Variable	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	$\hat{I}$
Year at marriage	74.88	5.973	0.074	Age at marriage	7.02	3.102	0.306
Year at birth	78.55	5.168	0.062				
Month of last birth	4210	171.300	0.039	Last closed birth	325.0	185.800	0.364
Month of next to last birth	1417	244.000	0.056	interval	0		
Date of marriage	10707	842.20	0.073	First birth interval		330.800	0.585
Month of first birth	9752	631.700	0.061		235.0 0		

The first of these variables is *age at first marriage*, which is derived directly from *year of marriage* (9) and *year of birth* (7). It is worth examining in some detail the way in which the observed value of $\hat{I}$ arises.

The value of $\hat{\sigma}_y^2$ for age at marriage is the variance of the difference between the `true' values of year of marriage and year of birth. Since these two variables are highly correlated, the value of $\hat{\sigma}_y^2$ for age at marriage is much lower than the value of $\hat{\sigma}_y^2$ for either of them.

This automatically implies that the denominator in $\hat{I}$ for age at marriage will be much smaller than that for two variables from which it is derived.

Similarly, the simple response variance for age at first marriage is the variance of the difference between the response deviations on the other two variables. In this case there are no *a priori* grounds for expecting a correlation between the two sets of response deviations.

If there were no correlation we would obtain a value of $\hat{\sigma}_\varepsilon^2$ of approximately 11.1. In fact, $\hat{\sigma}_\varepsilon^2$ is 3.1. This indicates that to some extent the reporting of age at marriage is of higher reliability than could be expected from the quality of the two dates from which it is derived. The value of $\hat{I}$ obtained, is, however, a valid measure of the reliability of the variable itself.

For the two birth interval variables the same general structure emerges. The values of $\hat{I}$ obtained are due to a dramatic reduction in $\hat{\sigma}_y^2$ counterbalanced in part by a decrease in $\hat{\sigma}_\varepsilon^2$. For both variables, but particularly for the first birth interval, the values of $\hat{I}$ are very large and indicate a high degree of unreliability.

The other two variables in table 6.8L are of a different type. *Number of children desired* is an attitudinal variable and thus might be expected to be particularly sensitive to the measurement process. Both this and *Ever-use of contraception* are variables for which the true value could change in the time period between the two interviews. Because of the form in which the data were collected, it is not possible to adjust these variables to take changes into account. Both show a high degree of inconsistency between the two interviews.

Figure 6.1L gives a diagrammatic presentation of the components of the simple total variance for the variables in table 6.8L.

Figure 6.1L about here

6.3.4 The range of values of $\hat{I}$ for Peru

Figure 6.1P gives the corresponding representation for 12 variables for Peru. As in Lesotho, the variables in figure 6.1P show a very wide range of values for $\hat{I}$ and the pattern is strikingly

similar. For the variable *Age* the data show a very high degree of reliability, with only 2 per cent of the simple total variance being attributable to simple response variance. At the other extreme, the variables *First birth interval* and *Number of children desired* show a high degree of unreliability, with values of I of 0.56 and 0.58 respectively. In the case of these two variables more than half the simple total variance is due to the response deviations - in other words, of the total variability in the responses less than half can be attributed to genuine differences in the underlying values of the variable in the population; the remainder is due to disturbances introduced into the observations by the measurement process itself.

Figure 6.1P about here

6.4 SRV for Different Subclasses of Respondents

The results in Figures 6.1L and 6.1P give an overall impression of the degree of reliability of responses for the key variables in the WFS. It is important to bear in mind, however, that most of the analysis of the data will be carried out on subsets of the whole sample, ie subclasses of the population. In section 6.4.1 we consider the variation in SRV across subclasses in Peru, concentrating on the extreme categories for age, education and urbanization. In section 6.4.2 we extend the analysis to Lesotho and consider in more detail the implications of this variation, concentrating on age and education subclasses.

6.4.1 SRV for subclass estimates for Peru

This section looks at three important sets of subclasses: age groups, education subclasses, and city size. The classification value of each of these is taken as reported in the first interview.

Table 6.10 gives the values of I for six selected subclasses - the youngest and oldest age groups; residents of Lima and rural residents; and, those with no education and those with seven or more years of education. The results in table 6.10 show the values of I for the extreme subclasses for each characteristic.

Table 6.10: *I* for six subclasses for Peru (values x 100)

Variable	Age < 25	Age ≥ 45	Lima	Rural	No educ.	Educ. ≥ 7 years	All
Age	06	29	01	03	04	01	02
Children ever born	02	03	01	03	03	01	02
Year of first birth	28	15	01	03	03	04	02
Age in 5 year groups	24	-	02	04	06	02	03
Year of last birth	09	03	01	06	04	02	03
Year of marriage	10	25	02	06	11	01	04
Marital duration	12	24	02	06	11	01	04
Education	06	06	08	16	-	-	06
Year of next to last birth	59	07	02	08	06	07	06
Births in past 5 years	05	28	05	13	11	06	09
Last closed birth interval	81	31	13	18	17	34	20
Age at marriage	16	29	11	36	40	07	20
Worked since marriage	31	40	21	43	53	21	30
Ever-use of contraception	30	36	34	62	45	37	35
First birth interval	57	66	29	77	82	42	56
No. of children desired	40	75	47	67	60	36	58

In general the results show a consistent pattern. For older women, for women with little education and for women residing in rural areas, the values of *I* are generally higher and in some cases much higher than for younger, better educated urban women. The background characteristics are not unrelated, of course; better educated women are generally younger and tend to live in urban areas. The number of cases in the sample (n = 1198) does not, however, permit good estimation of the degree of inconsistency for cells of a two-way or three-way classification.

A number of important conclusions emerge from the table. The value of *I* for the total sample

is not sufficient to evaluate all estimates based on a particular variable. In the case of *Age at marriage*, for instance, the value of I for the total sample is 0.20. In Lima the value is only 0.11 whereas in rural areas the value is 0.36. For the women with 7 or more years of education I is only 0.07 whereas for those with no education it is 0.40. Thus the degree of confidence we can have in the reliability of the data may be very different for different parts of the sample.

Figures 6.2 and 6.3 show in graphical form the partitioning of the simple total variance for the urbanization and age subclasses of two variables - *Marital duration* and *Number of children desired.* Marital duration is a relatively reliable variable - the overall variable of I is 0.04 - whereas number of children desired is extremely unreliable, $I = 0.58$.

Figures 6.2 and 6.3 about here

These figures indicate the implications and the pattern of the variation on reliability for different subclasses. The pattern is consistent - unreliability increases with age, rurality and lack of education. The first part of figure 6.3 has particular significance. It can be seen clearly from the figure that each and every age subclass has a higher value of I than the sample as a whole. This is not the case for any variable for any subclass other than age subclasses.

This occurs because the variable concerned (marital duration) is itself age-related. Since the sample as a whole is a cross-section of the population with ages ranging from 15 to 49, the variance of the true values $\left(\sigma_y^2\right)$ of the observations within a particular age subclass is much lower than the variance of the true values for the population as a whole. The response variance $\left(\sigma_\varepsilon^2\right)$, on the other hand, must be, on average across subclasses, equal to the response variance for the whole sample. This affects the three variables *Age, Marital duration* and *Year of marriage* very strongly, and is an important fact to bear in mind when considering the impact of response errors on analysis within age group.

Overall the analysis in this section indicates that the values of measures of reliability for the total sample provide only a rough guide to the reliability of results for subclasses of the population. This reservation is particularly note-worthy for the analysis of age-related variables, but is also relevant to analysis for any variables when dealing with subclasses containing a high proportion of rural, uneducated or older respondents.

6.4.2 SRV for subclass estimates for Lesotho

Building on the analysis of the Peru data, in this section the data from Lesotho are used to investigate in more detail two sets of subclasses that are relevant in almost all countries: education subclasses and age groups. Again the classification value in each case is taken as reported in the first interview.

Education subclasses

Table 6.11 presents the values of $\hat{\sigma}_y^2$ $\hat{\sigma}_\varepsilon^2$ and *1* for three education subclasses - those with 0-4 years of education, those with 5-6 years and those with seven or more years. In some ways it would have been preferable to use instead the categories no education, 1-5 years and six or more years, but the number of cases in the no education category would then have been too small to give reasonably stable results.

In evaluating the results in this table two different criteria can be considered. Since we are comparing the reliability across subclasses within each particular variable the simple response variance σ_ε^2 is in some ways the most appropriate measure. However, the implications of a particular value of σ_ε^2 in analyses involving more than one variable arise from the magnitude of the index of inconsistency *1*, which depends also on the variability of the true values in the subclass.

Table 6.11 Components of the simple total variance for the total sample and three education subclasses in Lesotho

Variable	Total sample (n = 609)			0-4 years' education (n = 191)			5-6 years' education (n = 253)			7 or more year' education (n = 165)		
	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	t	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	t	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	t	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	t
Year of last birth	29.0600	1.1180	0.037	43.5700	2.0580	0.043	23.1500	0.3384	0.014	18.3000	1.1230	0.058
Month of last birth	4210.0000	171.3000	0.039	6346.0000	322.0000	0.046	3347.0000	52.8800	0.015	2633.0000	162.9000	0.058
Year of next to last birth	28.7700	1.6490	0.054	42.5000	2.0100	0.043	22.8600	1.2750	0.049	18.9200	1.7460	0.084
Month of next to last birth	4147.0000	244.0000	0.056	6078.0000	302.0000	0.046	3337.0000	186.8000	0.049	2709.0000	257.8000	0.087
Year of first birth	67.7900	4.4560	0.062	68.2000	9.0590	0.113	66.8700	1.4140	0.021	56.6300	3.7050	0.061
Month of first birth	9752.0000	631.7000	0.061	9810.0000	1285.0000	0.112	9602.0000	195.2000	0.020	8189.0000	530.1000	0.061
Age	78.5500	5.1680	0.062	76.7200	8.7510	0.103	81.5100	3.4430	0.040	68.5500	3.2110	0.045
Age in five-year groups	3.1220	0.2325	0.070	3.0690	0.3717	0.109	3.1990	0.1626	0.048	2.6890	0.1470	0.052
Year of marriage	74.8800	5.9730	0.074	76.4500	10.4600	0.118	71.0200	3.6220	0.048	64.2400	4.1400	0.061
Marital duration	75.1800	5.9640	0.074	77.0200	10.4000	0.117	70.8800	3.7480	0.050	64.5900	4.0000	0.058
Children ever born	6.1800	0.5580	0.083	6.2430	0.8182	0.116	6.2900	0.3505	0.053	5.5330	0.5760	0.094
Years of education	4.2270	0.6460	0.131	2.7260	0.7925	0.223	0.2598	0.3332	0.571	0.4195	0.7870	0.652
Births in last five years	0.6875	0.1215	0.150	0.6720	0.1168	0.196	0.7111	0.0924	0.114	0.6660	0.1230	0.156
Age at first marriage	7.0170	3.1020	0.306	5.8570	4.1560	0.424	6.8970	3.1200	0.313	7.7550	1.7600	0.185
Last closed birth interval	324.9000	185.9000	0.364	410.0000	148.2000	0.255	3330.1000	233.0000	0.411	206.7000	155.3000	0.429
No of children desired	3.5230	3.2480	0.480	3.0470	4.6940	0.639	3.3020	2.5360	0.433	3.6800	2.8530	0.437
Ever-use of contraception	0.0695	0.0850	0.550	0.0510	0.0677	0.564	0.0703	0.0782	0.524	0.0880	0.1075	0.550
First birth	234.9000	330.8000	0.585	121.1000	628.2000	0.8564	335.8000	255.5000	0.462	244.4000	57.5200	0.191

The most striking feature of table 6.11 is that for all but two variables the value of $\hat{\sigma}_\varepsilon^2$ is largest for the lowest education group. This is consistent with our expectation that the quality of responses rises with level of education. The value of $\hat{\sigma}_\varepsilon^2$ for those with little education is typically between 40 and 100 per cent larger than $\hat{\sigma}_\varepsilon^2$ for the total sample.

In cases where the level of reliability overall is already low this suggests that there is even more cause for concern when the analysis is confined to this subclass. Two variables may be taken as examples. For the *First birth interval* the simple response variance for the total sample is estimated to be 331; for the low education subclass the value is 628, almost twice as large. For *Age at first marriage* the values are 3.1 and 4.2 respectively.

When the index of inconsistency is taken as the criterion the results follow the same pattern. This is reassuring since *I* depends not only on $\hat{\sigma}_\varepsilon^2$ but also on $\hat{\sigma}_y^2$, which is also an estimate. For twelve of the variables the value of *I* is largest for the lowest education subclass. Five variables for which this is not the case are the dates of the last and next to last births and the last closed birth interval, which is derived from them. The remaining variable, years of education, is a special case.

For the former set of twelve variables the value of *I* for the lowest education subclass is between 30 and 90 per cent higher than the value of *I* for the total sample. Generally the increase is due to the larger value of $\hat{\sigma}_\varepsilon^2$ for the subclass. Furthermore, the correlation between the response deviations for related variables carries through to each of the subclasses. Table 6.12 illustrates this for two sets of variables.

Table 6.12 Components of the simple total variance for two sets of related variables for Lesotho

Variable	Total sample (n = 609)			0-4 years' education (n = 191)			5-6 years' education (n = 253)			7 or more year' education (n = 165)		
	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	I	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	I	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	I	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	I
Month of first marriage	10783.000	860.000	0.074	11009.000	1506.000	0.118	10227.000	552.000	0.048	9251.000	596.000	0.061
Month of first birth	9752.000	631.700	0.061	9810.000	1285.000	0.112	9602.000	195.200	0.020	8189.000	530.100	0.061
Correlation between response deviations		0.780			0.770			0.640			0.950	
First birth interval	234.900	330.800	0.585	121.100	628.200	0.856	335.800	255.500	0.462	244.400	57.520	0.191
Age	78.550	5.168	0.062	76.720	8.751	0.103	81.510	3.443	0.040	68.550	3.211	0.045
Year of first marriage	74.880	5.973	0.074	76.450	10.460	0.118	71.020	3.622	0.048	64.240	4.140	0.061
Correlation between response deviations		0.720			0.780			0.560			0.760	
Age at first marriage	7.017	3.102	0.306	5.857	4.156	0.424	6.897	3.120	0.313	7.755	1.760	0.185

The first set of three variables shows the way in which the reliability of *Age at first marriage* is determined. The two basic variables are *Age* (or year of birth) and *Year of first marriage*. The difference between these is *Age at first marriage*. The reliability of measurement for *Age* and *Year of first marriage* can be expressed as $\hat{\sigma}_{\varepsilon}^2$. In both cases this is largest for the least educated subclass. The two other subclasses have similar values for $\hat{\sigma}_{\varepsilon}^2$. When we consider *Age at first marriage* we find that $\hat{\sigma}_{\varepsilon}^2$ for this variable is much lower than we would expect on the basis of the values for the two variables from which it is constructed. If the response deviations for the two component variables were uncorrelated, then the response variance for *Age at first marriage* would be equal to the sum of the response variances for the other two. It is reassuring to note that this is not the case. For all three subclasses there is a strong positive correlation between the response deviations for *Year of birth* and *Year of first marriage*. Thus respondents are making compensating errors in reporting year of birth and year of first marriage. The correlation between the response deviations for the two latter variables is between 0.6 and 0.8 in each case. The absolute size of the response variance is in fact less for *Age at first marriage* for all subclasses than for either of its component variables.

The implications for the reliability of reporting *Age at first marriage* are considerable. In the absence of the correlation between the response deviations between the two component variables the value of *I* would be about 0.6 for the total sample and between 0.5 and 0.8 for the subclasses. The actual values of *I* are 0.31 for the total sample and 0.19, 0.31 and 0.42 for the most, middle and least educated subclasses respectively. The reason that all the values of *I* are larger for *Age at first marriage* than for *Year of birth* and *Year of first marriage* is that the variance of the true values $\left(\hat{\sigma}_{y}^2\right)$ is much smaller for *Age at first marriage*.

The other set of three variables shows the same underlying pattern. The situation here

illustrates even more dramatically how unwise it is to assume anything about a composite variable on the basis of information about the component variables separately. The two component variables are *Month of first birth* and *Month of first marriage*. Both have low values of *1* - 0.07 for the total sample and between 0.02 and 0.12 for the subclasses. The two variables are used to calculate values of the *First birth interval*. There are two ways in which we can consider the reliability of measurement of the variables. The most basic measure is the simple response variance $\hat{\sigma}_\varepsilon^2$. In the case of all three variables $\hat{\sigma}_\varepsilon^2$ is largest for the least educated subclass - about twice the value for the total sample. Once again the response deviations for the two basic variables are highly correlated. There is a strong element of compensation in the errors in date of *First marriage* and date of *First birth*. In the absence of this correlation values of $\hat{\sigma}_\varepsilon^2$ of about 1500 for the total sample and 2800 for the least educated subclass might be expected. In fact, the values obtained were 331 and 628 respectively. This implies a correlation of about 0.8 between the two sets of response deviations.

The second measure of reliability is the index of inconsistency, *1*. This depends not only on $\hat{\sigma}_\varepsilon^2$ but also on $\hat{\sigma}_y^2$. The situation is similar to that described above for the first set of variables. For the total sample $\hat{\sigma}_y^2$ is about 11,000 for date of first marriage and 10,000 for date of first birth. The variation in the first birth interval is, of course, much less - only 235 for the total sample. Consequently the values of *1* are far greater for the first birth interval than for the other two variables despite the smaller values of $\hat{\sigma}_\varepsilon^2$, ranging from 0.19 for the most educated group to 0.86 for the least educated.

These two sets of variables illustrate some important principles in assessing measurement

error. First, the reliability - however measured - for a composite variable cannot be predicted in any uniform way from the reliabilities of the component variables. Secondly, there is a high degree of correlation between response deviations on related variables. Thirdly, the overall reliability of a variable is dependent on both the size of the measurement error and the extent of the variability among the true values for the variable concerned.

Table 6.13 presents the data for the deviant variables in table 6.11. The variables involved all relate to dates of more recent births. The two component variables are *Month of last birth* and *Month of next to last birth*. The composite variable is *Last closed birth interval*. The pattern of values of *I* across subclasses is inconsistent with that found for the other variables. The values of *I* for the least educated subclass are lower for all three variables than those for the most educated subclass.

The basic reason for this anomaly is the variation in the value of $\hat{\sigma}_y^2$ - the simple sampling variance, or the variation between the true values. The values of $\hat{\sigma}_y^2$ are largest for the least educated and smallest for the most educated, so that although the simple response variance is lower for the more educated respondents, the *relative* reliability is not.

The same strong positive correlation between response deviations illustrated in table 6.12 for the two component variables is present also in table 6.13. In this case this would have been expected (or at least hoped for), since the two variables are part of the birth history and are measured by an integrated set of questions in the questionnaire.

Table 6.13 Components of the simple total variance for a composite variable and its elements for Lesotho

Variable	Total sample (n = 609)			0-4 years' education (n = 191)			5-6 years' education (n = 253)			7 or more year' education (n = 165)		
	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	ρ	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	ρ	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	ρ	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	ρ
Month of last birth	4210.0	171.30	0.039	6346.0	322.00	0.046	3347.0	52.88	0.015	2633.0	162.90	0.058
Month of next to last birth	4147.0	244.00	0.056	6078.0	302.00	0.046	3337.0	186.80	0.049	2709.0	257.80	0.087
Correlation between response deviations		0.55			0.76			0.03			0.62	
Last closed birth interval	324.9	185.8	0.364	410.0	148.20	0.255	330.1	233.00	0.411	206.7	155.30	0.429

Table 6.14 Components of the simple total variance for the total sample and two age subclasses for Lesotho

Variable	Total sample (n = 609)			Age ≤ 25 years (n = 169)			Age ≥ 45 years (n = 64)		
	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	ρ	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	ρ	$\hat{\sigma}_y^2$	$\hat{\sigma}_\varepsilon^2$	ρ
Year of last birth	29.0600	1.1180	0.037	0.8935	0.3510	0.282	48.3900	1.6950	0.034
Month of last birth	4210.0000	171.3000	0.039	120.8000	61.2700	0.337	7043.0000	259.5000	0.035
Year of next to last birth	28.7700	1.6490	0.054	1.4600	1.7580	0.546	44.8900	1.4040	0.030
Month of next to last birth	4147.0000	244.0000	0.056	205.1000	269.0000	0.567	6368.0000	198.5000	0.030
Year of first birth	67.7900	4.4560	0.062	4.9300	2.0510	0.294	13.4000	5.8200	0.303
Month of first birth	9752.0000	631.7000	0.061	693.7000	283.1000	0.290	1949.0000	799.2000	0.291
Age	78.5500	5.1680	0.062	60.960	4.2580	0.411	2.1000	7.2550	0.776
Age in five-year groups	3.1220	0.2325	0.070	0.2040	0.1965	0.491	d.n.a.	d.n.a.	d.n.a.
Year of first marriage	74.8800	5.9730	0.074	6.0180	3.8100	0.388	9.4530	7.1350	0.430
Marital duration	75.1800	5.9640	0.074	5.8910	3.9770	0.400	9.5860	7.0210	0.423
Children ever born	6.1800	0.5580	0.083	1.0190	0.3160	0.237	7.3510	0.5950	0.075
Years of education	4.2270	0.6460	0.131	3.1900	0.7245	0.185	n.a.	n.a.	n.a.
Births in last five years	0.6875	0.1215	0.150	0.5370	0.1255	0.189	0.2970	0.0475	0.140
At at first marriage	7.0170	3.1020	0.306	2.8890	1.4160	0.329	7.2560	2.9090	0.286
Last closed birth interval	324.9000	185.8000	0.364	196.5000	64.4500	0.247	363.1000	296.4000	0.449
No of children desired	3.5230	3.2480	0.480	3.3830	2.1750	0.391	3.2480	3.0290	0.480
Ever-use of contraception	0.0695	0.0850	0.550	0.0500	0.0575	0.535	0.0200	0.0655	0.766
First birth interval	234.9000	330.8000	0.585	196.6000	74.2700	0.274	167.0000	592.8000	0.780

Age subclasses

Table 6.14 presents the values of *1* and of the components of the simple total variance for two age subclasses - those under 25 and those 45 and over. Age subclasses are widely used in the analysis of WFS data and it is important to assess the reliability of reporting for these subsets of the sample.

The most striking feature of table 6.14 is the contrast between the values of the simple response variance for the two subclasses. The older subclass provides substantially less reliable reporting than the younger one. The simple response variance for the older subclass is typically twice as large as that for the younger subclass, and in some cases the difference is even greater. A possible interpretation of this difference is that in recalling events, the extent of the unreliability is determined at least in part by the time elapsed between the event and the interview. This interpretation is supported by the internal evidence, which can be obtained by comparing the observed response variances for different events for the younger subclass. The simple response variance for year of last birth (a recent event) was 0.351; for year of next to last birth (a more distant event) 1.758; and for year of first birth (a still more distant event) 2.051.

An examination of the values of *1* provides significant evidence of the complex nature of the problem of evaluating response reliability. Overall, the pattern is that which could be expected. For ten of the seventeen variables the value of *1* is larger for the older group. Two examples of the expected pattern are *Desired family size* and *First birth interval*. The values of $\hat{\sigma}_y^2$ are approximately stable and the variation in the values of *1* is due to the greater unreliability of responses for the older group. This is essentially the pattern established in table 6.10 for the education subclasses. The pattern does not, however, hold for other variables.

The first major difference between the results of table 6.11 and those of table 6.14 can be illustrated by the variable *Year of first marriage*. The values of *1* for the younger and older

groups are 0.39 and 0.43 respectively. The two intermediate age groups (not given in table 6.13) have similar values. The value of *1* for the total sample is 0.07. Figure 6.4 gives a diagrammatic representation. The problem here is that the value of $\hat{\sigma}_y^2$ is very different for the subclasses and for the total sample. This is because of the restriction of a subclass to a particular age group necessarily reduces the possible variation in *Year of first marriage* considerably compared to the variation in the total sample of women aged 15-49. The value of $\hat{\sigma}_y^2$ for the total sample is 74.88; for the under 25 age group it is only 6.02; for the 45 and over age group it is 9.45. The same phenomenon occurs for *Age at first marriage*, *Marital duration, Age at first birth* and, of course, *Age*. These variables are all age-related, and when the subclasses are based on an age categorization the value of *1* for the total sample are no longer an `average' of the subclass values as they were for the education subclasses considered earlier. The alternative diagrammatic representation in figure 6.5 illustrates this.

Figure 6.4 and figure 6.5 here

The discussion above raises a fundamental question about the ranking of the variables in Figures 6.1P and 6.1L. The initial impression given by the *1* values is that *Age*, for example, is extremely reliably reported. Table 6.14 shows that this assessment is crucially dependent on the context in which the variable is used. When the total sample is being considered, or when a subclass is being used which is not age-related, the relative reliability of age reporting is high, as measured by the value of *1*. When the analysis is restricted to an age group, however, the situation changes and the measure of a relationship between age and any other variable is severely affected by the response deviations. The same caveat applies to *Marital duration*. It is worth noting that the contrast between the reliability of *Age at first marriage* and *Marital duration* and that of *Age* is transformed by this change of context. For the restricted subclasses the *1* values are all about 0.40, whereas for the total sample both *Age* and *Marital duration* have *1* values near 0.07. In fact, for the age subclasses *Age at first marriage* is more reliable than the other two variables.

The second major difference between the age and the education subclasses can be seen in the group of variables dealing with dates of children's births. For *Date of first birth* and *Date of last birth* the reliability of reporting, as measured by the simple response variance, is much higher for the younger age group. But because of the range of dates to which the measurement refers, the values of *I* are approximately equal for first birth and much higher for the younger age group for date of last birth. The same contrast holds for date of next to last birth.

Two other variables are worth considering. *Number of children desired* has essentially the same pattern as the date of birth variables and the values of *I* reflect this. *Births in the past five years* is different insofar as it is a variable whose main relevance is to the younger group. This is the only variable for which the simple response variance is substantially larger for the younger respondents.

Conclusion

The analysis in this section establishes a clear pattern of response variability across subclasses but also suggests that great care must be taken in using measures of reliability outside the context in which they are calculated. The quality of the data is highest for the younger and for the more educated respondents. This is not necessarily reflected in the values of the index of inconsistency, *I*, because of the dependence of *I* on the variation in the true values for the group of respondents under consideration. The most striking illustration of this is given by the analysis of age subclasses, where for age-related variables the values of *I* are all dramatically increased.

The other important finding is that there is a strong element of correlation between response deviations for related variables. Tables 6.12 and 6.13 give some examples of this and provide some reassurance about the quality of reporting of intervals.

Chapter 7 - FURTHER RESPONSE VARIANCE ISSUES IN WFS

This chapter addresses two particular issues. First, an analysis is presented of the relationship between interviewers' reported assessments of respondent reliability and cooperation and the simple response variance. Second, results are presented for the variance of the variance estimates themselves (the simple response variance, the index of inconsistency, and particular cases of the correlated response variance).

7.1 Interviewers' Assessment of Responses

At two stages during the course of the interview the interviewers were instructed to record their observations on an aspect of the respondent's replies to the question. Immediately after completing the birth history section of the questionnaire, and before asking the questions dealing with contraception, the interviewer was asked to tick one of the three boxes indicating the *reliability of the answers given in the birth history section*; the three categories given are GOOD, FAIR and POOR. The interviewer's instructions suggest guidelines for completing this question: if considerable probing was necessary for determination of the dates of birth and pregnancies, or if inconsistencies arose in the answers, or if the interviewer got the impression that the respondent was unsure of the answers, then the POOR box was appropriate; if the interviewer felt that the respondent was not telling the truth, then again the reliability was to be classified as POOR. In the opposite case, the reliability was to be classified as GOOD; in intermediate cases, involving a moderate amount of probing or correcting, the FAIR box was to be used. Once the interview had been completed the interviewer was asked to tick one of four boxes indicating the respondent's *degree of co-operation*; the four categories given were: POOR, FAIR, GOOD and VERY GOOD. The interviewer was instructed *not* to complete this section in the presence of the respondent.

In this subsection we look at the extent to which the interviewers' assessments of the respondents are reflected in the magnitudes of the response deviations. For this purpose we use the absolute value of the difference between the responses obtained from the two interviews for an individual as a measure of the response error. The magnitude is therefore

the difference in units (months, years, births, etc) between the responses at the first and second interviews. The response deviations themselves would be unsatisfactory since, by definition, they tend to cancel out over groups of individuals. The interviewer's assessments are taken from the first interview in each case.

7.1.1 Results for Lesotho

The results for the total matched sample of 609 cases in Lesotho are given in table 7.1L. A clear pattern emerges from the table. The magnitude of the response deviations are directly related to the interviewer's assessments. There are only three variables for which the differences are not statistically significant. In two of these cases the direction of the differences is in keeping with the general pattern (*Last closed birth interval* and *No of children desired*); the third case (*Ever-use of contraception*) is a special one, being a binary variable. The remaining fifteen variables all show statistically significant differences. The *Reliability* classification is slightly more successful in differentiating between the respondents on the birth history variables, as might be expected.

When the linearity component of the differences is tested separately (with one degree of freedom) the strength of the relationship is confirmed. For twelve of the variables in the case of the *Reliability* classification and ten in the case of the *Co-operation* classification the linearity component is significant at the 0.01 level.

It is interesting to note that both the reliability and co-operation assessments are effective in differentiating between respondents. Furthermore, the assessment of reliability, which is based on the responses in the birth history section, seems also to be relevant to the background variables such as age and age at marriage and even to the attitudinal question on number of children desired.

Table 7.1L Magnitude of response deviations cross-tabulated by interviewers' assessments

Variable	Reliability			Co-operation			Total
	GOOD	FAIR	POOR	GOOD/ VERY GOOD	FAIR	POOR	
Age	1.04	2.05	1.50	1.07	1.65	4.78	1.24
Children ever born	0.32	0.46	1.00	0.30	0.53	1.32	0.36
Year of first birth	0.73	1.62	5.00	0.77	1.48	4.58	0.94
Month of first birth	9.09	19.89	62.67	9.49	18.68	57.08	11.72
Age in five-year groups	0.20	0.38	0.12	0.21	0.28	0.80	0.23
Year of last birth	0.47	0.73	2.17	0.47	0.84	0.95	0.54
Month of last birth	5.97	9.43	26.83	5.97	11.02	10.79	6.85
Year of marriage	0.98	2.05	1.87	1.00	1.70	4.81	1.19
Marital duration	1.10	2.17	2.00	1.11	1.83	5.00	1.30
Years of education	0.49	0.80	1.25	0.50	0.82	1.19	0.56
Year of next to last birth	0.67	0.92	3.00	0.68	0.94	1.98	0.76
Month of next to last birth	8.49	11.64	39.33	8.57	12.19	26.64	9.59
Births in past five years	0.16	0.34	0.37	0.16	0.29	0.68	0.19
Last closed birth interval	7.50	8.27	15.42	7.80	7.15	10.29	7.74
Age at first marriage	1.16	1.63	2.00	1.18	1.44	3.03	1.26
Ever-use of contraception	0.20	0.13	0.25	0.19	0.20	0.00	0.19
First birth interval	7.31	14.06	66.96	7.87	10.29	64.20	9.33
No of children desired	1.38	1.77	1.57	1.42	1.65	1.63	1.46
Sample size	486	115	8	499	97	12	609

The number of individuals classified as POOR is small for both the criteria used by the interviewers - less than 2 per cent in each case - but the AVERAGE/FAIR category is also effective in identifying a group with high response variability.

The same analysis was carried out for the age subclasses and for the education subclasses described previously in chapter 6. Since the sample sizes are considerably smaller for the subclasses, the POOR group was amalgamated with the FAIR/AVERAGE group for the analysis. The pattern of results persisted for the subclasses, and the differences were statistically significant for the fertility variables despite the smaller sample sizes.

On balance, the results indicated that the interviewer's assessments are strongly related to the quality of the responses. There is, however, evidence of association between assessments and

education, age and place of residence of the respondent. It is not possible to determine completely the extent to which these are the characteristics on which the interviewers base their judgements, but the results for the subclasses suggest that the interviewers' assessments provide a useful further indicator of the quality of the responses.

It would appear that interviewers are reluctant to classify respondents as either POOR or FAIR on either criterion; almost 80 per cent of the respondents were classified as GOOD or better for each assessment. In the case of *Co-operation*, however, where two positive categories GOOD and VERY GOOD were provided, the interviewers were not particularly successful in differentiating between the two. This suggests that although there may be scope for extending the categorization used in the assessment of reliability, the naming of the categories requires further consideration.

7.1.2 Results for Peru

The results for the total matched sample of 1198 cases in Peru are given in table 7.1P. In the case of the *Co-operation* variable no assessment was available for 13 cases. These are excluded from the table.

The pattern of the results is clear here also, and is very similar to that for Lesotho. With the exception of only the two binary variables, *Whether worked* and *Ever-use*, there is a perfect *rank* correlation between the interviewer's assessments and the magnitudes of the response deviations. The differences between the categories are also statistically significant, the level of significance being less than 0.001 in 20 of the 24 comparisons. As with Lesotho both the reliability and co-operation assessments are effective in differentiating between respondents and again the assessment of reliability, which is based on the responses in the birth history section, seems also to be relevant to the background variables such as age and age at marriage and even to the attitudinal question on number of children desired.

Table 7.1P Magnitude of response deviations cross-tabulated by interviewers' assessments in Peru

Variable	Reliability			Co-operation			
	GOOD	FAIR	POOR	VERY GOOD	GOOD	FAIR	POOR
Age	0.58	0.95	1.31	0.30	0.65	1.05	1. 37
Children ever born	0.12	0.28	0.76	0.09	0.15	0.26	0.79
Year of first birth	0.45	0.98	1 .53	0.30	0.50	1.08	1.53
Age group	0.12	0.20	0.20	0.05	0.14	0.21	0.23
Year of last birth	0.30	0.64	0.65	0.16	0.37	0.54	1.26
Month of last birth	3.52	7.97	10.63	1.80	4.40	6.80	18.40
Year of marriage	0.98	1.84	2.33	0.71	1.21	1.56	2.37
Births in past 5 years	0.14	0.22	0.44	0.10	0.15	0.25	0.38
Last closed interval	6.30	10.90	14.50	5.10	7.20	9.10	23.00
Age at marriage	1.20	1.76	2.19	0.93	1.36	1.53	2.50
Whether worked	0.13	0.13	0.13	0.14	0.12	0.11	0.23
Ever-use of contraception	0.19	0.22	0.09	0.14	0.21	0.21	0.08
First birth interval	13.60	24.00	25.40	10.00	15.70	22.90	28.10
Desired no of children	1.14	1.47	1.88	0.96	1.25	1.46	2.16
Sample size	932	223	42	280	667	202	35

The number of individuals classified as POOR is small for both the criteria used by the interviewers - less than 4 per cent in each case - but the AVERAGE/FAIR category (REGULAR in the Spanish version of the questionnaire) is also effective in identifying a group with high response variability.

The same analysis was carried out for the urban and rural subclasses and for the five education subclasses described previously. Since the sample sizes are considerably smaller for the subclasses, the POOR group was amalgamated with the FAIR/AVERAGE group for the analysis. The pattern of results persisted for the urban and rural subclasses, and the differences were statistically significant for the fertility variables despite the smaller sample sizes. For all the education subclasses the same pattern emerged, although *reliability* differentiated better than *co-operation* in general. The results were least convincing in the lowest education group.

On balance, the results indicated that the interviewer's assessments are strongly related to the quality of the responses. As in Lesotho there is, however, evidence of association between assessments and education, age and place of residence of the respondent and it is not possible to determine the extent to which these are the characteristics on which the interviewers base their judgements; the results for the subclasses suggest that the interviewers' assessments provide a useful further indicator of the quality of the responses.

Again as in Lesotho it would appear that interviewers are reluctant to classify respondents as either POOR or FAIR on either criterion; more than 70 per cent of the respondents were classified as GOOD or better for each assessment.

In the case of *Co-operation*, however, where two positive categories GOOD and VERY GOOD were provided, the interviewers were quite successful in differentiating between the two; this contrasts with the case in Lesotho, where, in terms of the current analysis, the two categories did not discriminate. This suggests that there is scope for extending the categorization used in the assessment of reliability and that a wider choice, particularly of positive categories, might increase the usefulness of the indicator. The results for Lesotho indicate the importance of the precise words chosen for this purpose.

A note of caution may be appropriate here. Although the differences observed are large and of substantive significance, the proportion of the total variability in the response deviations which they explain is generally small.

7.2 Variance of the Variance Estimators

It has been emphasized throughout this report that the values of the measures presented in the tables of results are themselves estimates based on the observations in the sample. These values are subject to sampling variability and it is desirable that the magnitude of this variability should be estimated.

The procedure used in this section is the jackknife, first proposed as a method of reducing bias in ratio estimators and now widely used to estimate variances (see, for example, Kish and Frankel 1974, Kalton 1977). The basic steps are as follows:

(i) Divide the sample into a number k of random subgroups. These subgroups could be primary sampling units, or groups of primary sampling units.

(ii) Calculate the value of the measure, u say, leaving out each subgroup in turn. This will give a set of k values of the measure u. Denote these by u_{-1}, u_{-2}, ..., u_{-i}, ..., u_{-k} where u_{-i} is the value of u for the data ignoring subgroup i.

(iii) Calculate the pseudo-values (u_{*i}: $i = 1, ..., k$)
where $u_{*i} = ku - (k - 1)u_{-i}$

(iv) Calculate

$$u_* = \frac{1}{k} \sum_i u_{*i}$$

(v) The variance of u_* can be estimated by

$$var(u_*) = \frac{1}{k(k-1)} \sum_i (u_{*i} - u_*)^2$$

(vi) The estimated standard error of u_* is

$$se(u_*) = \sqrt{var(u_*)}$$

(vii) We use var(u_*) as an estimate of the variance of the measure u.

This procedure can be applied to measures based on the whole sample and also to measures based on subclasses.

7.2.1 The Simple Response Variance σ_ε^2

The full sample

One of the basic measures of response error used in this report is the simple response variance. Tables 7.2L and 7.2P present the estimates of σ_ε^2 , and the estimated variance, the estimated standard error and the estimated coefficient of variation of these estimates for the 18 variables previously considered for Lesotho and the 16 variables previously considered for Peru. The results in the tables are reassuring. The coefficient of variation of $\hat{\sigma}_\varepsilon^2$ is remarkably stable across variables, with most values close to 0.20 for Lesotho and 0.17 for Peru. The level of the values is satisfactory in that it provides reasonable confidence in the estimated values of σ_ε^2 . The range of values of cv $(\hat{\sigma}_\varepsilon^2)$ is from 0.10 to 0.27 in Lesotho and 0.07 to 0.28 for Peru, with the lowest values in both countries for *Births in the past five years* and *Ever-use of contraception.*

Table 7.2L $\hat{\sigma}_\varepsilon^2$ **, se** $\left(\hat{\sigma}_\varepsilon^2\right)$ **and cv** $\left(\hat{\sigma}_\varepsilon^2\right)$ **for the 18 variables in Lesotho**

Variable	$\hat{\sigma}_\varepsilon^2$	se $\left(\hat{\sigma}_\varepsilon^2\right)$	cv $\left(\hat{\sigma}_\varepsilon^2\right)$
Age	5.1680	1.30000	0.25
Children ever born	0.5580	0.1230	0.22
Year of first birth	4.4560	1.1880	0.27
Month of first birth	631.7000	163.2000	0.26
Age in five-year groups	0.2325	0.0520	0.22
Year of last birth	1.1180	0.1793	0.16
Month of last birth	171.3000	26.5800	0.16
Year of marriage	5.9730	1.2960	0.22
Marital duration	5.9640	1.3120	0.22
Education in years	0.6460	0.0787	0.12
Year of next to last birth	1.6490	0.2866	0.17
Month of next to last birth	244.0000	44.6900	0.18
Births in past five years	0.1215	0.0121	0.10
Last closed birth interval	185.8000	45.4300	0.24
Age at marriage	3.1020	0.5364	0.17
Ever-use of contraception	0.0850	0.0081	0.10
First birth interval	330.8000	80.9900	0.24
No of children desired	3.2480	0.6683	0.21

Table 7.2P $\hat{\sigma}_\varepsilon^2$, var $\left(\hat{\sigma}_\varepsilon^2\right)$, se $\left(\hat{\sigma}_\varepsilon^2\right)$, cv $\left(\hat{\sigma}_\varepsilon^2\right)$ **for the 16 variables in Peru**

Variable	$\hat{\sigma}_\varepsilon^2$	Var $\left(\hat{\sigma}_\varepsilon^2\right)$	se $\left(\hat{\sigma}_\varepsilon^2\right)$	cv $\left(\hat{\sigma}_\varepsilon^2\right)$
Age	1.1929	0.0335	0.1829	0.15
Children ever born	0.1485	0.0016	0.0395	0.27
Year of first birth	1.5746	0.1939	0.4404	0.28
Age in 5 year groups	0.0819	0.0001	0.0092	0.11
Year of last birth	0.6183	0.0121	0.1100	0.18
Year of marriage	3.0258	0.2833	0.5323	0.18
Marital duration	3.1238	0.2781	0.5274	0.17
Education	0.1270	0.0005	0.0220	0.17
Year of next to last birth	1.4970	0.0613	0.2476	0.17
Births in past 5 years	0.0897	0.0001	0.0076	0.08
Last closed birth interval	163.5427	1388.1861	37.2584	0.23
Age at marriage	3.4571	0.2924	0.5407	0.16
Worked since marriage	0.0664	0.0001	0.0091	0.14
Ever-use of contraception	0.0870	0.0000	0.0057	0.07
First birth interval	182.4751	1020.7824	31.9497	0.18
No. of children desired	2.2091	0.1919	0.4380	0.20

Subclasses

The estimates of σ_ε^2 in tables 7.2L and 7.2P are based on the whole samples for Lesotho and Peru (n = 609 and 1198 individuals, respectively). Many of the estimates used in practice and in WFS analysis are based on subclasses of the sample, where the number of individuals is much smaller. We would therefore expect the variance estimates to be less precise in these cases. Table 7.3L gives the results of the jackknife estimation of the variance for four important subclasses in Lesotho: respondents under 25; respondents over 45; respondents with 0-4 years of education; and those with more than seven years of education. The six variables presented are chosen to represent different levels of sensitivity to response errors. Table 7.3P presents the results of the jackknife estimation of the variance for five important subclasses for Peru for the same six variables: rural areas; metropolitan Lima; respondents

with no formal education; respondents under 25; and respondents over 45.

Table 7.3L $\hat{\sigma}_\varepsilon^2$ **, se** $\left(\hat{\sigma}_\varepsilon^2\right)$ **and cv** $\left(\hat{\sigma}_\varepsilon^2\right)$ **for six variables for four subclasses for Lesotho**

Variable	Children ever born			Year of last birth		
Subclass	$\hat{\sigma}_\varepsilon^2$	se $\left(\hat{\sigma}_\varepsilon^2\right)$	cv $\left(\hat{\sigma}_\varepsilon^2\right)$	$\hat{\sigma}_\varepsilon^2$	se $\left(\hat{\sigma}_\varepsilon^2\right)$	cv $\left(\hat{\sigma}_\varepsilon^2\right)$
Under 25 yr	0.3160	0.1255	0.40	0.351	0.1483	0.42
Over 45 yr	0.5950	0.3834	0.64	1.695	0.8796	0.52
Educ. 0-4 yr	0.8182	0.2506	0.31	2.058	0.4439	0.22
Educ. 7+ yr	0.5760	0.1769	0.31	1.123	0.4143	0.37
All	0.5580	0.1230	0.22	1.118	0.1793	0.16
Variable	Age at marriage			Ever-use of contraception		
Subclass	$\hat{\sigma}_\varepsilon^2$	se $\left(\hat{\sigma}_\varepsilon^2\right)$	cv $\left(\hat{\sigma}_\varepsilon^2\right)$	$\hat{\sigma}_\varepsilon^2$	se $\left(\hat{\sigma}_\varepsilon^2\right)$	cv $\left(\hat{\sigma}_\varepsilon^2\right)$
Under 25 yr	1.416	0.3550	0.25	0.0575	0.0111	0.19
Over 45 yr	2.909	1.0728	0.37	0.0655	0.0296	0.45
Educ. 0-4 yr	4.156	1.2680	0.31	0.0677	0.0131	0.19
Educ. 7+ yr	1.760	0.4060	0.23	0.1075	0.0165	0.15
All	3.102	0.5364	0.17	0.0850	0.0081	0.10

Table 7.3P $\hat{\sigma}_\epsilon^2$, se $\left(\hat{\sigma}_\epsilon^2\right)$ and cv $\left(\hat{\sigma}_\epsilon^2\right)$ for six variables for five subclasses for Peru

Variable	Children ever born			Year of last birth			Marital duration		
Subclass	$\hat{\sigma}_\epsilon^2$	se $\left(\hat{\sigma}_\epsilon^2\right)$	cv $\left(\hat{\sigma}_\epsilon^2\right)$	$\hat{\sigma}_\epsilon^2$	se $\left(\hat{\sigma}_\epsilon^2\right)$	cv $\left(\hat{\sigma}_\epsilon^2\right)$	$\hat{\sigma}_\epsilon^2$	se $\left(\hat{\sigma}_\epsilon^2\right)$	cv $\left(\hat{\sigma}_\epsilon^2\right)$
Rural	0.2397	0.1252	0.52	0.8097	0.3070	0.38	4.8137	1.0742	0.22
Lima	0.0429	0.0110	0.56	0.3830	0.2024	0.53	1.4974	0.3437	0.23
No education	0.2471	0.1275	0.52	0.8325	0.1230	0.15	6.6647	1.9418	0.29
Under 25	0.0257	0.0081	0.32	0.1019	0.0396	0.39	0.8473	0.1797	0.21
Over 45	0.3176	0.1890	0.59	0.9285	0.2757	0.30	5.3354	1.1533	0.22
All	0.1485	0.0395	0.27	0.6183	0.1100	0.18	3.1238	0.5274	0.17
Variable	Age at marriage			Ever-use of contraception			First birth interval		
Subclass	$\hat{\sigma}_\epsilon^2$	se $\left(\hat{\sigma}_\epsilon^2\right)$	cv $\left(\hat{\sigma}_\epsilon^2\right)$	$\hat{\sigma}_\epsilon^2$	se $\left(\hat{\sigma}_\epsilon^2\right)$	cv $\left(\hat{\sigma}_\epsilon^2\right)$	$\hat{\sigma}_\epsilon^2$	se $\left(\hat{\sigma}_\epsilon^2\right)$	cv $\left(\hat{\sigma}_\epsilon^2\right)$
Rural	5.4189	1.1915	0.22	0.0881	0.0134	0.15	309.6	70.4	0.23
Lima	1.9583	0.5118	0.26	0.0627	0.0104	0.17	96.2	20.5	0.21
No education	6.9489	1.7698	0.25	0.0775	0.0156	0.20	369.8	121.1	0.33
Under 25	0.8628	0.1460	0.17	0.0749	0.0164	0.22	87.9	25.1	0.29
Over 45	6.4465	1.5927	0.25	0.0803	0.0147	0.18	374.4	187.4	0.50
All	3.4571	0.5407	0.16	0.0870	0.0057	0.07	182.5	31.9	0.17

An interesting feature of tables 7.3L and 7.3P is the variation in the values of the simple response variance, σ_ε^2 , across subclasses. For five of the six variables (the exception is *Ever-use of contraception*) the simple response variance is much larger for the less educated and over 45 subclasses than for the under 25 and more educated subclasses; for Peru the SRV is also larger for rural than for the urban subclass. This is in keeping with the results previously discussed in section 6.4, and serves as a reminder of the need for caution in extending the results for the total sample to particular subclasses of interest.

The second point about tables 7.3L and 7.3P is that the coefficient of variation σ_ε^2 is in general larger for the subclasses than for the total sample. This is not surprising as the estimate of σ_ε^2 is based on fewer observations in the case of subclasses than in the case of the total sample, and consequently the variance (or the standard error) of σ_ε^2 might be expected to be correspondingly larger.

What is perhaps worth noting is that the coefficient of variation of σ_ε^2 is much more stable across subclasses than the simple response variance itself. This is particularly noticeable in the case of *Marital duration, Age at marriage* and *First birth interval.* In fact this is reassuring since it conforms to the theoretical expectation for the variance of a variance estimator of this kind.

In general, if $\hat{\sigma}^2$ is an estimator of σ^2 based on $n - 1$ degrees of freedom, then

$$var(\hat{\sigma}^2) = \frac{\mu_4 - \mu_2^2}{n} + \frac{2}{n(n-1)}\mu_2^2$$
$$= \sigma^2\left[\frac{\beta_2 - 1}{n} + \frac{2}{n(n-1)}\right]$$

where $\beta_2 = \mu_4/\sigma^4$ and μ_2, μ_4 are the second and fourth moments for the parent distribution.

For large n,

$$var(\hat{\sigma}^2) \doteq \frac{(\beta_2 - 1)\sigma^4}{n}$$

and

$$se(\hat{\sigma}^2) \doteq \sqrt{\frac{\beta_2 - 1}{n}.\sigma^2}$$

Consequently

$$cv(\hat{\sigma}^2) \doteq \sqrt{\frac{\beta_2 - 1}{n}}\ \alpha\frac{1}{\sqrt{n}}$$

Thus on theoretical statistical grounds we would expect the ratio of coefficient of variation of $\hat{\sigma}_\varepsilon^2$ for the total sample to that for a subclass to be approximately inversely proportional to $(n_t/n_c)^{1/2}$, where n_t and n_c are the sample sizes for the total sample and the subclass respectively.

For the subclasses in table 7.3L this would imply ratios of 1.9:3.1:1.8:1.9:1 for the coefficients of variation for the under 25, over 45, less educated, more educated and total sample respectively. These are remarkably close to the ratios found in table 7.3L. Equally satisfying is the fact that even for the subclasses, the coefficients of variation are reasonably small. Except for the smallest subclass (respondents over 45), the coefficients of variation are of the order of 0.2-0.3; for the smallest subclass they are of the order of 0.5.

For Peru, the derivation implies that the ratio of the $cv(\sigma^2)$ for the subclasses to the $cv(\sigma^2)$ for the total sample should be between 1.5:1 and 2.75:1, similar to the ratios found in table 7.3P. The coefficients of variation for the subclasses are also of the order of 0.2 to 0.3 here, except for *Children ever born* and *Year of last birth* for which they are larger and less stable. This is understandable since these are the variables with the lowest degree of response variance.

7.2.2 The index of inconsistency, *I*

The index of inconsistency, I (defined by $\sigma_\varepsilon^2/(\sigma_y^2 + \sigma_\varepsilon^2)$), measures the proportion of the simple total variance which is due to the simple response variance. The estimates $\hat{I}$ of I obtained from the data are used extensively in sections 6.3 and 6.4 to describe the sensitivity of variables to response errors. In table 6.8L and figures 6.1L and 6.1P the values of $\hat{I}$ for the total sample are presented, while tables 6.10 through 6.14 and figures 6.2 through 6.5 give the values of $\hat{I}$ for major subclasses. The validity of the conclusions drawn from these tables and figures depend on the precision of the estimates of $\hat{I}$.

Lesotho

Table 7.4L presents the results of the jackknife estimation of the variance of $\hat{I}$ for the six variables and four subclasses considered for Lesotho in section 7.2.1 above. The variables span the range of observed values of $\hat{I}$ and the subclasses represent the extremes of the characteristics considered.

The pattern of variation in the values of $\hat{I}$ is similar to that for $\hat{\sigma}_\varepsilon^2$. The values of $\hat{I}$ for the variables presented are: for *Children ever born*, 0.08; for *Year of last birth*, 0.04; for *Marital duration*, 0.07; for *Age at marriage*, 0.31; for *Ever-use of contraception*, 0.55; and for *First birth interval*, 0.58.

Table 7.4L *f*, se(*f*) and cv (*f*) for six variables and four subclasses

Variable	Children ever born			Year of last birth			Marital duration		
Subclass	*f*	se(*f*)	cv(*f*)	*f*	se(*f*)	cv(*f*)	*f*	se(*f*)	cv(*f*)
Under 25 yr	0.237	0.0844	0.36	0.282	0.0701	0.25	0.400	0.0820	0.20
Over 45 yr	0.075	0.0422	0.56	0.034	0.0247	0.73	0.423	0.0851	0.20
Educ. 0-4 yr	0.116	0.0319	0.27	0.043	0.0144	0.33	0.117	0.0336	0.29
Educ. 7+ yr	0.094	0.0360	0.38	0.058	0.0283	0.66	0.058	0.0198	0.34
All	0.083	0.0179	0.22	0.037	0.0072	0.19	0.074	0.0177	0.34
Variable	Age at marriage			Ever-use of contraception			First birth interval		
Subclass	*f*	se(*f*)	cv(*f*)	*f*	se(*f*)	cv(*f*)	*f*	se(*f*)	cv(*f*)
Under 25 yr	0.329	0.0935	0.28	0.535	0.0974	0.18	0.274	0.1065	0.39
Over 45 yr	0.286	0.1737	0.61	0.766	0.2205	0.29	0.780	0.1726	0.22
Educ. 0-4 yr	0.424	0.0755	0.18	0.564	0.0812	0.14	0.856	0.1178	0.14
Educ. 7+ yr	0.185	0.0433	0.23	0.550	0.0719	0.13	0.191	0.0884	0.46
All	0.306	0.0373	0.12	0.550	0.0521	0.09	0.585	0.0919	0.16

The coefficients of variation of the estimates of I for the total sample are similar to those for the corresponding values for $\hat{\sigma}_\varepsilon^2$. For the subclasses the pattern is also similar, with the smallest subclass (respondents over 45) having the largest coefficient of variation for $\bar{I}$ in the case of four of the six variables. Only five of the 24 coefficients presented exceed 0.4; the average value for the others is about 0.23. These are comparable to the corresponding values for $\hat{\sigma}_\varepsilon^2$, and justify some confidence of the $\bar{I}$ values for subclasses. Three examples are given below. These are differences commented on in the text of section 6.4 in chapter 6.

In general, the variance of the difference between two random variables x_1 and x_2 is

$$\text{var}(x_1 - x_2) = \text{var}(x_1) + \text{var}(x_2) - 2\text{cov}(x_1, x_2).$$

For the differences discussed here, the model for the simple response variance implies that the covariance term is zero. Hence,

$$\text{var}(\bar{I}_1 - \bar{I}_2) = \text{var}(\bar{I}_1) + \text{var}(\bar{I}_2).$$

Table 7.5LA gives the computations for those comparisons of values of $\bar{I}$. The last two columns give the difference in $\bar{I}$ for the subclasses and the estimated standard error of this difference.

For the three contrasts given in table 7.5LA the estimated precision of the estimated difference is sufficiently high to warrant the conclusion that there is a real difference in the values of the index of inconsistency in these cases. It would be unwise, however, to have too much faith in the absolute value of the difference estimated. If we were justified in constructing a normal 95 per cent confidence interval for the difference in $\bar{I}$ between women with little education and those with more than seven years' education for the first birth interval, the confidence interval would be 0.665 ± 0.294 or 0.371, 0.959.

Table 7.5LA Standard errors of contrasts of I for subclass pairs for Lesotho

Variance	Subclass	I	se(I)	$I_1 - I_2$	se($I_1 - I_2$)
First birth interval	0-4yr education	0.856	0.1178	0.665	0.147
	7+yr education	0.191	0.0884		
First birth interval	Age under 25	0.274	0.1065	0.506	0.203
	Age over 45	0.780	0.1726		
Age at marriage	0-4yr education	0.424	0.0755	0.239	0.087
	7+yr education	0.185	0.0433		

Furthermore, not all the apparent differences in values of I are estimated precisely enough to justify much confidence. An example is given in table 7.5LB.

Table 7.5LB A counter-example to table 7.5LA

Variance	Subclass	I	se(I)	$I_1 - I_2$	se($I_1 - I_2$)
Ever-use of contraception	Age under 25	0.535	0.0974	0.231	0.241
	Age over 45	0.766	0.2205		

The difference in values of I is 0.23. The estimated variance for the difference, however, suggests that this apparent difference may result simply from the sampling variance of the estimates involved. The estimated standard error is greater than the estimated difference, and thus a 95 per cent normal confidence interval would be:

$$0.231 \pm 0.472 \text{ or } (-0.241, 0.703).$$

This does not mean that there is no difference between the values of the index of inconsistency

for the two subclasses for *Ever-use of contraception.* It does mean, however, that additional evidence would be necessary before the presence of the difference could be established beyond reasonable doubt.

Peru

Table 7.4P presents the results of the jackknife estimation of the variance of $\bar{I}$ for the six variables and five subclasses considered for Peru in section 7.2.1 above. As with Lesotho the variables span the range of observed values of $\bar{I}$ and the subclasses represent the extremes of the characteristics considered.

The pattern of the variation in the values of $\bar{I}$ is similar to that for σ_ε^2 . The variables are

arranged in order of increasing $\bar{I}$ overall - for *Children ever born* $\bar{I}$ is 0.02; for *Year of last birth*, 0.03; for *Marital duration*, 0.04; for *Age at marriage*, 0.20; for *Ever-use of contraception*, 0.35; and for *First birth interval*, 0.56.

Table 7.4P *I*, se (*I*) and cv (*I*) for six variables for five subclasses for Peru

Variable	Children ever born			Year of last birth			Marital duration		
Subclass	*I*	se (*I*)	cv (*I*)	*I*	se (*I*)	cv (*I*)	*I*	se (*I*)	cv (*I*)
Rural	0.023	0.0156	0.68	0.056	0.0220	0.39	0.066	0.0110	0.17
Lima	0.005	0.0017	0.32	0.129	0.0067	0.52	0.020	0.0044	0.22
No education	0.023	0.0145	0.63	0.040	0.0098	0.25	0.115	0.0347	0.30
Under 25	0.013	0.0057	0.43	0.081	0.0353	0.44	0.116	0.0293	0.25
Over 45	0.275	0.0164	0.60	0.029	0.0098	0.33	0.236	0.0569	0.24
Variable	Age at marriage			Ever-use of contraception			First birth interval		
Subclass	*I*	se (*I*)	cv (*I*)	*I*	se (*I*)	cv (*I*)	*I*	se (*I*)	cv (*I*)
Rural	0.363	0.081	0.22	0.542	0.145	0.27	0.763	0.084	0.11
Lima	0.109	0.030	0.28	0.329	0.059	0.18	0.289	0.071	0.25
No education	0.402	0.074	0.18	0.434	0.107	0.25	0.833	0.139	0.17
Under 25	0.157	0.034	0.22	0.304	0.066	0.22	0.569	0.086	0.15
Over 45	0.267	0.092	0.34	0.372	0.076	0.20	0.672	0.227	0.34

The coefficients of variation for the estimates of I for the subclasses are of the same order of magnitude as those for σ_ε^2 . The least stable estimates are those for the first two variables - the level and variability of the *cv*'s are high for these variables. For the remaining four variables the situation is more satisfactory. The range of the *cv*'s is 0.11 to 0.34 with an average value near 0.22. These are comparable to the corresponding values for σ_ε^2 , and justify some confidence in the conclusions reached on the basis of a comparison of the $\hat{I}$ values for subclasses. Three examples, commented on in the text of section 6.4, are given below. Table 7.5PA gives the computations for those comparisons of values of $\hat{I}$. The last two columns give the differences in $\hat{I}$ for the subclasses and the estimated standard error of this difference.

Table 7.5PA Standard errors of contrasts of $\hat{I}$ values for subclass pairs in Peru

Variable	Subclass	$\hat{I}$	se ($\hat{I}$)	var ($\hat{I}$)	var ($\hat{I}_1$-$\hat{I}_2$)	se ($\hat{I}_1$-$\hat{I}_2$)	$\hat{I}_1$-$\hat{I}_2$
Age at marriage	Rural	0.363	0.081	0.00659	0.00750	0.087	0.254
	Lima	0.109	0.030	0.00091			
Age at marriage	No education	0.402	0.074	0.00546	0.00570	0.075	0.332
	7+ years	0.070	0.015	0.00024			
First birth interval	Rural	0.763	0.084	0.00713	0.01211	0.110	0.474
	Lima	0.289	0.071	0.00498			

For the three contrasts given in table 7.5PA the estimated precision of the estimated difference is sufficiently high to warrant the conclusion that there is a real difference in the values of the index of inconsistency in these cases. It would be unwise however to have too much faith in the absolute value of the difference estimated. If we were justified in constructing a normal 95 per cent confidence interval for the difference in I between women in rural areas and those in Lima for the first birth interval, the confidence interval would be 0.474 ± 0.216 or (0.258, 0.690).

Furthermore, not all the apparent differences in values of $\bar{I}$ are estimated precisely enough to justify much confidence. An example is given in table 7.5PB.

Table 7.5PB A counter-example to table 7.5PA

Variable	Subclass	$\bar{I}$	se ($\bar{I}$)	var ($\bar{I}$)	var ($\bar{I}_1$-$\bar{I}_2$)	se ($\bar{I}_1$-$\bar{I}_2$)	$\bar{I}_1$-$\bar{I}_2$
Marital duration	Under 25	0.116	0.029	0.00086	0.00409	0.064	0.120
	Over 45	0.236	0.057	0.00323			

The difference in the values of $\bar{I}$ is 0.12; in fact the estimate of I for the youngest subclass is less than half that for the oldest subclass. The estimated variance for the difference, however, suggests that this apparent difference may result simply from the sampling variance of the estimates involved. The estimated standard error is equal to more than half the estimated difference, and thus a 95 per cent normal confidence interval would be:

$$0.120 \pm 0.125 \text{ or } (-0.005, 0.245).$$

This does not mean that there is no difference between the values of the index of inconsistency for the two subclasses for *Marital duration.* It does mean, however, that additional evidence would be necessary before the presence of the difference can be established beyond reasonable doubt. In the case of this particular difference the consistency of the pattern of response variance across age subclasses is so marked that the contrast for a particular variable receives support from the contrast for other variables. Indeed the only exceptions to the pattern of variation in $\bar{I}$ are explicable in terms of the variation in the values of the simple sampling variance in these cases (in some age-related subclasses, for instance).

7.2.3 The intra-interviewer correlation coefficient, ρ_i

The technical problems caused by the disruption of the fieldwork execution in Peru complicated the estimation of the correlated response variance and made the estimation of

precision impracticable. There is also a problem in estimating the correlated response components for cross-classes in that as the sample size decreases the estimates themselves become subject to larger variances. The data from Lesotho however did allow some scope for this and variance estimates for the correlated response variance were also derived using the jackknife.

The variable with the largest correlated interviewer variance for Lesotho was *Ever-use of contraception*. Table 7.6 gives the estimated values of ρ_i for the full sample and for five subclasses for this variable, together with their estimated standard errors. The computations are for the data from the main survey.

Table 7.6 Estimated values of ρ_i for cross-classes for *Ever-use of contraception* in Lesotho

	Total sample	Age < 25	Age 25-34	Age 35-44	Educ. 1-5yr	Educ. 6yr +
ρ_i	0.084	0.061	0.121	0.072	0.088	0.052
$se(\rho_i)$	0.018	0.042	0.038	0.044	0.034	0.024
M_s	1.00	0.31	0.33	0.27	0.44	0.48

From the table it can be seen that the values of ρ_i for the subclasses are consistent with the assumption that ρ_i remains constant across subclasses. In no case is the value of ρ_i more than one standard from the value of 0.084 obtained for the total sample. This evidence was presented in section 5.3 of chapter 5 to support the model used there.

7.2.4 Discussion

All the measures of response variability presented in this report are estimates based on the sample of respondents observed in the main survey and the re-interview survey, and are thus

themselves subject to sampling variance. In this section the three basic measures of response variability are considered - the simple response variance σ_ε^2 , the index of inconsistency I, and the intra-interviewer correlation coefficient ρ_i. The procedure used to estimate the variance of the estimates is the jackknife, a general method applicable to any measure.

The results are encouraging and indicate that the precision of the estimates is sufficiently high to justify statements about the general level of response errors and to confirm broad patterns of variation across variables and across subclasses. It is clear from the computations however that not all apparent differences in level are sufficiently supported by the evidence - tables 5LB and 5PB provide examples.

The model presented in section 5.3 of chapter 5 describes the total variance of an estimate in terms of five factors: the simple total variance; the synthetic intracluster correlation coefficient for the sample design; the synthetic intracluster correlation coefficient for the fieldwork design; the average cluster take; and the average interviewer workload size. The model is analogous to that generally used to describe the total sampling variance. A number of important assumptions are made in the model. First, it is assumed that the cross-classes are uniformly distributed across clusters and interviewers; in the context of WFS surveys, age subclasses are likely to satisfy this condition at least approximately. Secondly, it is assumed that the intra-cluster and intra-interviewer correlation coefficients remain constant for cross-classes. The evidence on this is less convincing, although it seems a useful approximation in practice. In particular, the evidence elsewhere for the intra-cluster correlation coefficient suggests that it is reasonably stable (see Verma et al, 1980, for instance). The data in table 7.6 in relation to the intra-interviewer correlation are encouraging also, though further investigation of the behaviour of the intra-interviewer correlation would be desirable.

Chapter 8 THE UNDERLYING RELIABILITY (QUALITY) OF BINARY DATA: THE UNITED STATES CURRENT POPULATION SURVEY (CPS)

8.1 Introduction

In this chapter we will consider the particular issue of response variance for a binary survey variable and the estimation of this variance where reinterview data are available. We wish to estimate the reliability of the responses and to understand what influences the magnitude of this reliability. An important feature of the work is to distinguish between the measurement error itself and problems which arise in its estimation.

As a starting point for the analysis we consider the standard survey model of response error and the estimation and interpretation issues which arise with this model. The results presented here are based primarily on O'Muircheartaigh (1986). That paper used data from the U.S. Current Population Survey (CPS) to estimate some of the characteristics of the measurement errors and from these derived some conclusions about the measurement process, but also identified some deficiencies in these conclusions.

In chapter 9 we will consider an alternative conceptualization of the measurement error using a log-linear model relating the errors to the characteristics of the measurement situation. We will reconsider the estimation problems in the light of the new model and reanalyze the CPS data in order to see whether the new approach helps to illuminate the underlying measurement process.

The overall structure of models of response error is described in chapter 2. The standard model of response variance developed at the U.S. Bureau of the Census (Hansen, Hurwitz and Bershad, 1961) is given in section 2.3.2.

The observation y_{jt} is the response obtained for individual j at trial t. The survey is assumed to be conceptually repeatable and t will take the values 1, 2, ... etc for repetitions of the survey. The observation is considered to consist of a true value (y_j) for individual j and a

response deviation which may be partitioned into a bias or systematic distortion (β_j) for individual j, and a variable response error (ε_{jt}) that varies from trial to trial. We have for each individual j two observations y_{j1} and y_{j2}

The data consist of the original interview and a reinterview for a set of individuals. Both interviews may be considered to be conducted under the same (or similar) essential survey conditions, and there is therefore no possibility of assessing the bias terms $\{b_j\}$ from the data.

Where y is a binary variable (such as "employed, not employed") the data from the two sets of interviews can be written as in figure 8.1. Under the same assumptions as before, it can be shown that $(b+c)/n$, which is called the gross difference rate (*gdr*), is equal to twice the estimate of the simple response variance.

Figure 8.1 An interview-reinterview table.

		Original Interview		
		$y_{j1} = 1$	$y_{j2} = 1$	
Reinterview	$y_{j2} = 1$	a	b	a+b
	$y_{j2} = 0$	c	d	c+d
		a+c	b+d	n

The *gdr* is the proportion of individuals from whom different responses were obtained on the two occasions and has a strong intuitive appeal as a measure of reliability. There are intuitive deficiencies in the standard mathematical formulation in the case of binary variables; in particular the error term ε_{jt} cannot take a spread of values for a particular element in the population. It is true that in most cases we would not consider this form to be a plausible measure of response variance. It is nevertheless worth noting that the SRV even in this case is extremely plausible when expressed in the form of the *gdr*.

The data on which the analysis in this chapter is based come from the CPS reinterview program. The CPS is a monthly household sample survey conducted by the U.S. Bureau of the Census which provides estimates of labor force characteristics; in a given month each individual is classified according to labor force status. Any eligible respondent can report for all the individuals in the household. Each household is in the CPS sample for four months, then leaves the sample for eight months, and returns to the sample for a further four months. Thus at any given time the sample will contain households which will have been in the sample for 1, 2, ..., 8 months. As part of the quality control procedures, a subsample of about 1 in 18 (about 5,000) households is reinterviewed in the same month. At the time of the original interview the interviewer does not know that the household will be reinterviewed. The person conducting the reinterview (the reinterviewer) is never the original interviewer, and is usually a senior interviewer or supervisor.

For 80 percent[1] of the reinterview sample a reconciliation between the original and reinterview responses is carried out as part of the reinterview program. The reinterviewers for these cases have access to the original responses but are instructed not to consult them until after the reinterview has been completed. The reconciliation itself is carried out on a separate form. The term "contaminated" is used to describe these cases; the data used are the reinterview responses before the reconciliation operation is carried out. For the remaining 20 percent of the sample, no reconciliation is carried out and the reinterviewers have no access to the original responses; these are called the uncontaminated cases. The data consist of the interview-reinterview cases from January 1982 through January 1984 - a total of 126,122 cases. For some purposes the data for February 1984 through December 1984 are also used - a further 55,000 cases.

People are not asked whether they are employed or unemployed, nor do interviewers classify their respondents' employment status. Instead, respondents are asked a series of questions about their activities and, on the basis of their replies, are classified for the survey by

1
In 1988 the CPS has changed to a 75%-25% split of reconciled to unreconciled reinterview (see chapter 15 in Biemer et al.(1991) for details).

"Employment Status Recodes". The most important classifications are "Employed" and "Unemployed". These two together comprise the "Civilian Labour Force " (CLF). Some subclassifications of the "Employed" group are "Non-agricultural", "Full-time" and "Part-time".

The reinterview program for the CPS has been in place for forty years. In a paper in 1964, Hansen, Hurwitz and Pritzker presented the gross difference rates (*gdr*) for these six important labour force categories (each treated as a binary variable) for the period 1958-61. The *gdr* is expressed as a percentage in the tables. In table 8.1, the *gdr* for the six variables are given also for the 1982-84 period. The comparison is confined to the unreconciled cases.

Table 8.1: *gdr* (per cent) for six variables, 1958-61 and 1982-84

	Unreconciled data	
Variable	1958-61	1982-84
CLF	4.1	5.0
Employed	3.9	3.8
Unemployed	1.9	3.1
Non-agric.	3.5	3.8
Full-time	4.4	4.4
Part-time	4.6	5.0

The table suggests only modest changes in simple response variance since 1958. For the variables "employed" and "full-time" there is no change; for "non-agricultural", "part-time" and "in civilian labour force" there is possibly a slight increase; and for "unemployed" a more substantial increase.

The overall values of the *gdr* (=2SRV) given in table 8.1 are affected by a large number of factors not all of which are directly related to data quality. In section 8.2 these factors are briefly described. The subsequent sections examine some of the factors more intensively and in the final section some recommendations are made about the ways in which the data evaluation program for the CPS might be modified and how the data quality of the CPS might

eventually be improved.

8.2 Factors Affecting the Estimated Simple Response Variance

In order to understand the process which produces the estimates of simple response variance given in table 8.1 it is necessary to consider the factors which are likely to affect the values of the estimates. A diagrammatic representation is given in figure 8.2. The most important distinction is that between the two sets of factors on the left, which affect the response variance itself, and the set of factors on the right, which affect not the response variance but its estimation. The estimates presented in table 8.1 (and in later tables) are affected by all three sets of factors.

Figure 8.2 about here

If two independent observations were obtained for each question for each individual under identical essential survey conditions, then the estimated SRV would be an unbiased estimate of the actual SRV for the survey. Even in this case however, it would be unwise to consider the SRV to be a single indivisible quantity.

Because of the way in which the data are collected for the CPS we may expect the *quality of reporting* to differ for different individuals. First, the data for some individuals will be based on self-reports and the data for other individuals will be based on reports by others (proxy reports). There is no prior justification for assuming that the reliability of reporting will be the same for these two situations. Second, the characteristics of the respondents (for instance, age, sex, position in the household) will vary and these variations may be related either to the quality of their self-reports or of their proxy-reports. There may also be interactions between the characteristics of the respondents (those doing the reporting) and of the subjects (those being reported on) in the case of proxy reports. Third, some of the interviews are conducted face-to-face (personal interviews) and others by telephone; this may also affect the response variance. Fourth, the reliability of the data may be related to the number of months for which the household has already been in the sample; there is evidence (Bailar (1975) on rotation

group bias, for instance) that data quality is related to month in sample.

This set of factors must be considered in relation to the inherent measurement problems for the survey variables. Although a labour force characteristic such as whether a person is employed, unemployed or out of the labour force may seem to be a factual item, and one that can easily be verified, there are people for whom labour force status is an attitude that cannot be verified from records (see Bailar and Rothwell (1984) for a discussion of this issue). There also may be interactions between the characteristics of the respondents and the characteristics which influence the relevance of the questions and the stability of the true values for particular individuals.

The factors described above relate to the magnitude of the actual response variance of the survey estimates. The set of factors on the right of figure 8.2, however, are factors which affect the estimate of response variance but do not affect the response variance itself. These factors represent departures from the ideal (or assumed) situation of independent reenumeration under identical essential survey conditions. Each of them may invalidate to a greater or lesser extent the estimates given in table 8.1. Each of these factors involves design issues in the reinterview program.

First, the households selected for reinterview are not an equal probability sample from the population and therefore the estimation formula should be modified to take this into account. Second, the reinterviews are not an independent replication of the original interviews. There are some procedural differences between the original interviews and the reinterviews: the interviewers for the reinterview tend to be more senior than the generality of interviewers, and the distribution of mode of interview (telephone vs. personal) may also differ. These factors invalidate (or at least weaken) the assumption that the reliability of responses is the same in the two interviews. In addition, there are factors which create a dependence between the response deviations in the two interviews.

For instance, if the same individual responds to the questions on the first and second interviews, there is a chance that s/he will give the same response on the two occasions simply

because s/he remembers the response given on the first occasion. This would imply a positive correlation between the response deviation on the first interview (ε_{j1}) and that on the second interview (ε_{j2}). The availability to the reinterviewer of the responses to the first interview could have a similar effect for the contaminated 80 percent of the sample, as could communication between members of the household when different individuals respond on the two occasions. The effect might differ for different self/proxy combinations, and could be expected to differ for different time intervals between the first and second interviews.

These design and estimation issues are considered in section 8.3. In section 8.4 the reporting issues are examined and in section 8.5 the two are combined in an attempt to understand the factors influencing data reliability.

8.3 Estimation Issues

Some of the problems that arise in the estimation of simple response variance apply generally to reinterview programs but the CPS procedures have particular strengths and weaknesses in this context. First, contaminating 80 percent of the reinterviews by allowing the reinterviewer to have access to the original responses at the time of the reinterview reduces the value of these cases in the estimation of reliability. Second, any eligible respondent can report for the whole household in either interview; thus some observations are self-reports and others are proxy-reports. In terms of the interview-reinterview data being used in the estimation procedure, five categories of self/proxy combinations may usefully be identified -- these are given in figure 8.3.

Figure 8.3: Self/proxy combinations in the CPS

Original interview	Reinterivew	Code	Estimate
self	self	P_0	$s^2_{\varepsilon_0}$
proxy	self	P_1	$s^2_{\varepsilon_1}$
self	proxy	P_2	$s^2_{\varepsilon_2}$
proxy	same proxy	P_4	$s^2_{\varepsilon_4}$
proxy	different proxy	P_5	$s^2_{\varepsilon_5}$

Table 8.2: *gdr* by type of reporting: unreconciled and contaminated

	Unreconciled					
Variable	P_0	P_1	P_2	P_4	P_5	All
CLF	3.58	5.78	5.70	4.73	12.03	4.93
Employed	2.84	4.82	4.19	3.39	9.51	3.76
Unemployed	2.27	4.08	3.06	3.10	6.43	3.08
Non-agric.	2.72	4.65	4.36	3.46	9.09	3.72
Full-time	3.63	5.69	3.75	3.75	7.48	4.33
Part-time	4.05	5.99	4.43	4.43	10.28	5.01

	Contaminated					
Variable	P_0	P_1	P_2	P_4	P_5	All
CLF	1.96	3.58	2.69	2.38	4.97	2.52
Employed	1.47	2.92	2.25	1.72	4.10	1.93
Unemployed	1.13	2.27	1.72	1.49	2.74	1.53
Non-agric.	1.41	2.94	2.34	1.76	3.88	1.92
Full-time	2.13	4.01	3.32	2.19	3.56	2.53
Part-time	2.17	4.27	3.31	2.27	4.73	2.67

Table 8.2 presents the estimates of the *gdr* for each of the five self/proxy combinations for both the unreconciled (20 percent) and contaminated (80 percent) data for the six labour force variables previously considered in table 8.1 for the period January 1982 - January 1984. The slight discrepancies between the values for the category "All" and those in table 8.1 arise from the small proportion of cases for which a self/proxy categorization was not possible.

The pattern of variation exhibited by the results in table 8. 2 is informative. First, it is clear that the contamination has a marked effect on the estimates of response variance -- values for the unreconciled cases are about twice as large as those for the contaminated cases. Second, the self/proxy categories also exhibit considerable variation, and the pattern of differences is similar for each of the six variables. Overall the lowest values for the *gdr* are found in the P_0 (self/self) category, and these are uniformly lower than the values for the P_4 (proxy/same proxy) category. The two mixed self/proxy categories P_1 and P_2 have similar values, with a tendency for P_1 to be greater than P_2, especially for the contaminated data. The values for P_5 are very large relative to the other four categories.

Each of the columns in table 8.2 represents a particular combination of interview/reinterview respondent mix and reinterview type. We may postulate that the particular combination of characteristics will be related to the degree to which the interview and reinterview observations may be assumed to be independent. The comparative lack of independence may provide information on the effect of particular aspects of the reinterview program design, and should make it possible to improve the estimates of simple response variance.

In the remainder of this section the data in table 8.2 will be used to estimate the between-trial correlation for the response deviations, to investigate the effect of contamination, and to examine the possible impact of within-household communication on the estimates of simple response variance.

8.3.1 Between-trial correlation

The general estimate of the simple response variance, s_ε^2 , is

$$s_\varepsilon^2 = \frac{1}{2n_{jes}} \Sigma \left(\varepsilon_{j1} - \varepsilon_{j2}\right)^2 \tag{8.1}$$

If the analysis is confined to the unreconciled cases there are two identifiable factors in table 8.2 which may affect the size of the estimated SRV. First, whether an observation is a self-report or a proxy-report; and second, whether the two observations for an individual were obtained from the same respondent on the two occasions. When the same respondent reports on the two occasions it is possible that s/he will recall the answer given in the first interview and may tend as a result of this to give the same answer in the reinterview even if the first answer has been incorrect. If so, this could lead to a correlation, ρ, between the response deviations in the two observations.

Such a correlation between trials would produce an underestimation bias in the estimates of the SRV. The expected value of s_ε^2 would then become

$$E\left(s_\varepsilon^2\right) = \frac{1}{2}\left(\sigma_{\varepsilon_1}^2 + \sigma_{\varepsilon_2}^2 - 2\rho\sigma_{\varepsilon_1}\sigma_{\varepsilon_2}\right) \tag{8.2}$$

Under the assumption that $\sigma_{\varepsilon_1}^2 = \sigma_{\varepsilon_2}^2$, this becomes

$$\begin{aligned} E\left(s_\varepsilon^2\right) &= \frac{1}{2}\left(2\sigma_\varepsilon^2 - 2\rho\sigma_\varepsilon^2\right) \\ &= \sigma_\varepsilon^2\left(1-\rho\right) \end{aligned} \tag{8.3}$$

which would imply that for a correlation of 0.3, for instance, the SRV would be underestimated by 30 percent.

Previous work

A number of attempts have been made to estimate the value of ρ. Bailar (1968) in the context of the CPS considered the effect of increasing the time interval between interviews as a means of investigating ρ but did not produce specific estimates of the effect.

O'Muircheartaigh (1984a) in the context of the World Fertility Survey used a fortuitous aspect of the fieldwork to estimate the effect indirectly. In the case of Peru, due to the long period over which the fieldwork extended, it was possible to compare the magnitudes of estimated response deviations for varying time intervals between the interview and the re-interview; for the rural sector in particular there were time intervals ranging from 1 month to 10 months. A simple regression analysis, taking the squared response deviation as the dependent variable, was carried out for the set of variables. Statistically significant results were obtained for four variables - *Children ever born, Age group, Births in past five years,* and *Month of last birth.* Assuming that the effect of recalling the first response has disappeared after 10 months, the results indicated that the estimates of σ_ε^2 should be inflated by a factor of 1.3 approximately.

An examination of the residuals from the simple linear regression suggested that for some variables a more appropriate model would include a quadratic term in the time interval. For *First birth, Age at marriage* and *Year of marriage* this modified model produced a good fit, but the indicated under-estimation was only about 10 per cent in each case.

The evidence provided by this analysis is not by any means conclusive, particularly since, although an attempt was made to exclude real changes over time from the data, there was some difficulty in achieving this for *Births in the past five years.* Refinement of the estimation procedure and evidence from other surveys would be necessary before any substantial modification of the estimated σ_ε^2 should be introduced. Bailar's (1968) data suggested that the effect is negligible for many items, which is the situation we found for nine out of the 13 variables considered.

The current analysis

The combinations of interview/reinterview respondents for the CPS data set provide an opportunity to estimate the value of ρ directly. If the simple response variance for self-reports is denoted by $\sigma^2_{\varepsilon(s)}$ and that for proxy-reports by $\sigma^2_{\varepsilon(p)}$, the expected value of the estimated SRV for the columns of table 8.2 are as follows:

Table 8.3: Expected values of SRV for different interview-reinterview respondent combinations

Category	Code	Estimate	Expected value
self/self	P_0	$s^2_{\varepsilon_0}$	$\sigma^2_{\varepsilon(s)}\ (1-\rho)$
proxy/self	P_1	$s^2_{\varepsilon_1}$	$\left[\sigma^2_{\varepsilon(p)} + \sigma^2_{\varepsilon(s)}\right]/2$
self/proxy	P_2	$s^2_{\varepsilon_2}$	$\left[\sigma^2_{\varepsilon(s)} + \sigma^2_{\varepsilon(p)}\right]/2$
proxy/same proxy	P_4	$s^2_{\varepsilon_4}$	$\sigma^2_{\varepsilon(p)}\ (1-\rho)$
proxy/different proxy	P_5	$s^2_{\varepsilon_5}$	$\sigma^2_{\varepsilon(p)}$

These expected values provide two possible methods of estimating ρ:

(i) Using P_0, P_4 and P_1, P_2 gives an estimate $\hat{\rho}_1$

$$\hat{\rho}_1 = \frac{\left(s_{\varepsilon_1}^2 + s_{\varepsilon_2}^2\right) - \left(s_{\varepsilon_0}^2 + s_{\varepsilon_4}^2\right)}{\left(s_{\varepsilon_1}^2 + s_{\varepsilon_2}^2\right)} \tag{8.4}$$

which provides the following estimates of ρ for the six variables:

CLF	Emp.	Unemp.	Non-agric.	Full-time	Part-time
0.28	0.31	0.25	0.31	0.37	0.34

These values, which are all close to 0.3, suggest that for all of the cases where the same respondent provides the information on the two occasions the usual estimate of SRV is biased downwards by about 30 percent.

(ii) Using P_4 and P_5 only gives an alternative estimate $\hat{\rho}_2$

$$\hat{\rho}_2 = \frac{s_{\varepsilon_5}^2 - s_{\varepsilon_4}^2}{s_{\varepsilon_5}^2} \tag{8.5}$$

which provides the following estimates of ρ:

CLF	Emp.	Unemp.	Non-agric.	Full-time	Part-time
0.61	0.64	0.52	0.62	0.50	0.57

The values of $\hat{\rho}_2$ are all higher than the corresponding values of $\hat{\rho}_1$ and imply that the estimates of SRV for the categories P_0 and P_4 are biased downwards by more than 50 percent.

To consider the impact of ρ on the estimated value of the SRV for the whole sample, denote by π_i the proportion of cases belonging to the category P_i. Then the expected value of s_ε^2 is

(from table 8.3)

$$E\left(s_\varepsilon^2\right) = \pi_0\sigma^2_{\varepsilon(s)}\,(1-\rho) + \pi_1\left(\sigma^2_{\varepsilon(s)} + \sigma^2_{\varepsilon(p)}\right)/2 + \pi_2\left(\sigma^2_{\varepsilon(s)} + \sigma^2_{\varepsilon(p)}\right)/2 + \pi_4\sigma^2_{\varepsilon(p)}\,(1-\rho) + \pi_5\sigma^2_{\varepsilon(p)}$$

The correct value of s_ε^2 for the survey is

$$\sigma^2_\varepsilon = \pi_0\sigma^2_{\varepsilon(s)} + \pi_1\left(\sigma^2_{\varepsilon(s)} + \sigma^2_{\varepsilon(p)}\right)/2 + \pi_2\left(\sigma^2_{\varepsilon(s)} + \sigma^2_{\varepsilon(p)}\right)/2 + \pi_4\sigma^2_{\varepsilon(p)} + \pi_5\sigma^2_{\varepsilon(p)}$$

Thus the bias in s_ε^2 is

$$\rho\left[\pi_0\sigma^2_{\varepsilon(s)} + \pi_4\sigma^2_{\varepsilon(p)}\right] \tag{8.6}$$

The first row of table 8.4 gives the unadjusted estimates of the *gdr* for the unreconciled data. The second and third rows present adjusted estimates using the values of $\hat{\rho}_1$ and $\hat{\rho}_2$ respectively.

Table 8.4: Unadjusted and adjusted estimates of *gdr* (= 2SRV)

	Variable					
Estimate	CLF	Emp.	Unemp.	Non-agric.	Full-time	Part-time
Unadjusted	4.97	3.81	3.08	3.76	4.37	5.03
Adjusted ($\hat{\rho}_1$)	6.15	4.85	3.74	4.83	6.00	6.64
Unadjusted ($\hat{\rho}_2$)	7.58	5.98	4.46	5.89	6.56	7.76

Figure 8.4 about here

Two conclusions emerge from these results. First, the evidence is strong that there is a non-negligible between-trial correlation and that the usual estimate significantly underestimates the SRV. The lower estimates (based on $\hat{\rho}_1$) assume that the cases in categories P_1 and P_2 are free

from correlation. To the extent that there is communication within the household between the two respondents, the estimates $\hat{\rho}_1$ will be too low, and thus it may be reasonable to assume that the estimates adjusted using $\hat{\rho}_1$ are conservative. Second, it is clear from the divergence between the estimates $\hat{\rho}_1$ and $\hat{\rho}_2$ that the simple model is inadequate. No parsimonious model will ever represent the data perfectly, but it is clear that this model ignores some important terms. One possibility which is intuitively reasonable is to allow the between-trial correlation to differ for self-reports and proxy-reports. This unfortunately does not help, as the data will not produce valid estimates of the two correlations jointly (i.e. values with $|\rho| \leq 1$). This suggests, as does an inspection of table 8.2, that other characteristics of the respondents besides their self/proxy status must be introduced into the model. This is done in section 8.4.

8.3.2 Communication and contamination

The unreconciled cases represent the closest approximation in the data to independent reinterviews. It is possible to use the ratio of the *gdr*'s for contaminated and unreconciled cases to assess the impact of contamination on the estimate of SRV. The final column in table 8.5 gives these ratios for the whole sample; the values for all the variables are close to 0.5, indicating that the effect of allowing the reinterviewer to have access to the original responses is to reduce the estimate of SRV by 50 percent.

This reduction has nothing to do with the quality of the original data. It results from the impact that access to the original responses has on the responses recorded for the reinterview. The reinterviewer instructions stipulate that the original responses should <u>not</u> be used during the reinterview but are to be used only afterwards to reconcile discrepancies between the two sets of responses. It is clear that this is not what happens in practice. This result is not new and was first pointed out by Hansen, Hurwitz and Pritzker (1964). What is perhaps of additional interest is the way in which the ratios vary from one column to another in table 8.5.

Table 8.5: Ratio of contaminated to unreconciled *gdr* for different self/proxy combinations

Variable	P_0 self/self	P_1 proxy/self	P_2 self/proxy	P_4 proxy/proxy	P_5 proxy/ different proxy	All
CLF	0.56	0.63	0.48	0.50	0.42	0.51
Employed	0.53	0.59	0.53	0.50	0.43	0.51
Unemployed	0.50	0.56	0.56	0.48	0.43	0.50
Non-agric.	0.53	0.63	0.53	0.50	0.43	0.52
Full-time	0.59	0.71	0.56	0.59	0.48	0.58
Part-time	0.53	0.71	0.50	0.50	0.45	0.53

The two principal categories P_0 and P_4 show a broadly similar set of values; these are the categories in which the same respondent reported in both interviews. Where different respondents provided the responses in the two interviews (categories P_1, P_2 and P_5) the effect of contamination differed according to the type and order of respondents.

The values in table 8.5 for the category P_5 are the lowest of all. The effect of contamination ranges from 52 percent to a 58 percent reduction in the estimate of variance. The P_5 category is an outlier in many respects. The SRV for this category is very high and P_5 seems to be the source of the difficulty in obtaining a single estimate of the between-trial correlation. This issue is examined further in section 8.4.

According to the assumptions of the model (table 8.3) the situations represented by P_1 and P_2 are identical; in both cases one response is a self-report and the other is a proxy-report -- the only difference is in the order. There is a substantial difference, however, between the effects of contamination in two cases: for P_1 the reduction in the estimated variance ranges from 29 percent to 44 percent; for P_2 the reduction is between 44 percent and 52 percent. This suggests that the extent to which the reinterviewer is affected by the original responses depends on the source of the original response.

In comparing the estimated response variance for P_1 and P_2 one possibility which has not yet

been considered is that the order of respondents may affect the difference (or similarity) between the two responses even in the absence of contamination. Table 8.6A presents the values of the *gdr* for the unreconciled cases for P_1 and P_2 and also the ratio of the two. P_1 comprises the cases where the proxy-report is the original response and the self-report is the reinterview response. In these cases even if the original (proxy) respondent discusses the response with the subject (the individual about whom the report is made) it is unlikely that this will affect the subject's self-report in the reinterview. If, on the other hand, the self-report occurs first -- as is the case with P_2 -- and the original (self) respondent discusses the response with the second (proxy) respondent, it is much more likely that this communication will affect the proxy-report in the reinterview. Thus the values of the *gdr* can be expected to be somewhat higher (and closer to the true value) for P_1 than for P_2, and the difference may be attributed to the effect of communication within the household.

Table 8.6A: The effect of communication: *gdr* for P_1 and P_2 for unreconciled data

Variable	P_1 proxy/self	P_2 self/proxy	P_1/P_2
CLF	5.78	5.70	1.01
Employed	4.82	4.19	1.15
Unemployed	4.08	3.06	1.33
Non-agric.	4.65	4.36	1.07
Full-time	5.69	6.04	0.94
Part-time	6.00	6.78	0.89

There is some evidence from the data that the order of the respondents affects the estimate of response variance, particularly for the variables "employed" and "unemployed" although it should be emphasized that the values of *gdr* are estimates and that the ratios P_1/P_2 are also subject to sampling variation.

For the contaminated data the values of *gdr* for P_1 and P_2 may also be expected to differ. given that access to the original responses affects the estimate of response variance it is reasonable to hypothesize that the source of the original response may have an additional impact on the way in which the reinterviewer obtains or records the second response. The

ratios in the final column of table 8.6B support this hypothesis. The values for P_1 are substantially higher than the values for P_2 for all six variables. This suggests that when the first response is a self-report the reinterviewer allows this to influence the response (or the recording of the response) for the reinterview. when the first response is a proxy-report the effect of the first response on the second response (which is a self-report) is much less.

The ratios in table 8.6B need to be considered in conjunction with the ratios in table 8.6A -- it is the difference between the ratios (or the ratio of the ratios) that represents the additional contamination effect for P_2 cases over that for P_1 cases.

Table 8.6B: Additional contamination for P_2 cases: P_1 and P_2 for contaminated data

Variable	P_1 proxy/self	P_2 self/proxy	P_1/P_2
CLF	3.58	2.69	1.33
Employed	2.92	2.25	1.30
Unemployed	2.27	2.72	1.32
Non-agric.	2.94	2.34	1.26
Full-time	4.01	3.32	1.21
Part-time	4.27	3.31	1.29

8.3.3 Other design issues

The design of the reinterview sample does not give equal probabilities of selection to each individual, whereas the estimate of SRV is the appropriate estimate for the equal probability case. It is easy to show that this will lead to potential bias in the estimate and that the size of the bias will depend on the variability among the selection probabilities and the relationship between the selection probabilities and the response variances for different individuals. Empirical investigation indicated that for the CPS data examined the effect is negligible, although it would be preferable to modify the estimate so that it would be robust even if the empirical relationship between the selection probabilities and the response variances were to change.

The time interval between the two interviews also varies considerably, and can be more than two weeks in some cases. However analysis of response variance by time interval showed no consistent pattern or relationship. The time intervals involved are all short compared to those examined in other studies.

Some other design factors are not included explicitly in the analysis. In particular, the differences between the original interview and the reinterview procedures are not taken into account. To the extent that the reinterviews are better (in the sense of being less variable) than the original (i.e. the main survey) interviews, the estimates of response variance will be underestimates. If the reinterviews are better (in the sense of being less biased) then the estimates of response variance will include a component corresponding to this change in bias. The data set does not however contain information to permit the separate identification of these components.

8.3.4 Summary

This section has considered two important effects of the design of the reinterview program on the estimates of response variance. Using the unreconciled data, estimates of the between-trial correlation of the response deviations were obtained which demonstrated that the existing (usual) estimate of response variance is seriously biased downwards. Table 8.4 gives the unadjusted and adjusted estimates.

Two different estimates of the correlation were obtained depending on which equations in the model (table 8.3) were used. Further evidence of the inadequacy of the model was obtained from table 8.5 which compared the contaminated and unreconciled data for different respondent combinations. This provided the basis for the examination in section 3.2 of the extent of communication within household (table 8.6A) and the differential effect of contamination for different respondent orderings (table 8.6B).

Although the results in this section are useful, the combination of (i) the incompatibility of the two estimates of between-trial correlation and (ii) the evidence of differential communication

and contamination for different respondent combinations suggests that the model being used fails to specify important factors which relate to response variables. In the following section some of these issues are considered.

8.4 Reporting Issues

The characteristics of the respondents in a survey have been shown to be related to the quality of the responses obtained. For fertility data, for instance, O'Muircheartaigh (1984a, 1984b) showed that respondent age and education were related to the size of the simple response variance. In addition to these extra-role characteristics of the respondent the self/proxy response status may be important, as may the relationship between the respondent and the subject. In this section these issues are explored for the CPS.

8.4.1 Type of Respondent

For CPS data the identity of the respondent is recorded in the record for each subject (the subject being the individual about whom the survey variables are being reported). Although the respondent's *characteristics* are not included on the records for the subjects for whom s/he responds it is possible, though not easy, to obtain the information from the respondent's own subject record. This was done for the full data set, though in a small proportion of cases information on the respondent for one or other interview was not available.

Three factors are considered here -- the relationship of the individual to the head of household (*kin*), the age of the individual (*age*) and the mode of interview/reinterview (*mode*). The definitions of the categories are given in figure 8.4.

Figure 8.4

Relationship to head of household

Kin = 1	Head of household with relatives in household
Kin = 2	Head of household without relatives in household
Kin = 3	Wife of head of household
Kin = 4	Other relative

Age

Age = 2	24 years and under
Age = 3	25 - 34
Age = 4	35 - 44
Age = 5	45 - 54
Age = 6	55 - 64
Age = 7	65 years and over

Mode of interview

Mode = 1	Telephone (original)	Telephone (reinterview)
Mode = 2	Personal (original)	Telephone (reinterview)
Mode = 3	Personal (original)	Personal (reinterview)
Mode = 4	Telephone (original)	Personal (reinterview)
Mode = 5	Mixed (telephone and personal) in either interview	

It is important to emphasize at this stage that the choice of respondent is not random. The interviewer may choose as respondent any eligible member of the household and the probability of being selected as respondent is determined by the judgement of the interviewer, availability of the individual and many other unmeasured factors. This qualification applies also to the other tables in this chapter.

In table 8.7 the values of the *gdr* estimates are presented for the four principal categories of relationship to head of household. In order to avoid the effect of interactions between respondent and subject characteristics, and possible differential effects of communication and contamination, the data are confined to the unreconciled cases and to self-reports.

Table 8.7: *gdr* by relationship to head of household for self-reports (unreconciled data)

Variable	Kin 1	Kin 2	Kin 3	Kin 4	All
CLF	3.55	2.47	3.89	8.16	3.58
Employed	2.85	2.28	3.07	4.08	2.84
Unemployed	2.94	1.85	1.72	7.48	2.27
Non-agric.	2.66	2.32	2.80	4.76	2.72
Full-time	3.97	2.71	3.65	6.46	3.63
Part-time	3.74	3.22	4.34	7.82	4.05

"Heads of household without relatives in the household" (kin 2) have the lowest *gdr* in general, although "wives of heads of household" (kin 3) have an exceptionally low *gdr* for "unemployed". "Heads of household" (kin 1) and kin 3 have similar levels of response variance. The most noteworthy group is the "other relatives of head of household" (kin 4) group. The simple response variance for this group is generally twice as large as for any other group, and it particularly bad for the categories "unemployed" and "CLF".

The results in table 8.8 provide a further indication of the way in which reliability of reporting varies by type of respondent; the table again is confined to unreconciled self-reports. There is strong evidence from the table that consistency is higher for older respondents in the CPS. The results for "unemployed" are those that show the clearest pattern. The estimated *gdr* decreases from a level of 5.02 for the youngest age group (24 and under) to a level of 0.54 for the oldest group (65 and over) and there is a monotonic trend between these extremes. For the variables generally, the highest variability occurs for the younger respondents and the lowest for the older respondents.

Table 8.8: *gdr* by age group for self-reports (unreconciled data)

Variable	Age 2	Age 3	Age 4	Age 5	Age 6	Age 7
CLF	5.81	3.04	4.20	3.71	2.88	3.12
Employed	3.17	2.43	3.34	3.38	2.40	2.69
Unemployed	5.02	3.04	2.72	2.26	1.28	0.54
Non-agric.	3.57	2.57	3.21	2.90	2.32	2.26
Full-time	5.02	3.88	4.27	5.16	3.20	1.51
Part-time	5.42	4.58	4.64	4.35	3.12	2.80

Further analyses were carried out to investigate differences in reliability for other categories of respondents. Differences were found by gender and race of respondent, for instance. The general conclusion to be drawn from the results is that many characteristics of the respondent are related to the quality of reporting. The importance of these findings depends, however, on the extent to which a choice between respondents with different characteristics is possible in the field. In a household with two 38 year olds, for example, it is not useful to know that reporting is more reliable for 65 year olds.

Before turning to a consideration of applications of the results above, one further comparison is presented here. Much of the interviewing in the CPS is done by telephone (from the interviewer's home) although the choice between personal (face-to-face) and telephone interviewing is not randomized. When examining estimates of the response variance the cases must be classified by the combination of mode of interview on the two occasions. This is done in table 8.9. In this case all of the cases are included in the table.

Table 8.9: *gdr* by mode of interview/reinterview (*mode*)

	Variable						
Mode	CLF	Employed	Unemployed	Non-agric	Full-time	Part-time	No of cases
1 tel/tel	2.70	2.14	1.60	2.10	2.77	3.07	78,490
2 pers/tel	3.06	2.36	1.71	2.40	2.91	3.13	32,772
3 pers/pers	5.04	3.02	4.38	2.99	3.39	3.28	7,287
4 tel/pers	5.62	3.94	3.45	3.81	3.63	3.89	2,260
5 mixed/mixed	3.30	2.75	1.83	2.75	3.52	3.41	4,691
All	3.01	2.30	1.83	2.28	2.89	3.13	125,900

Since an even higher proportion of the reinterviews is conducted by telephone, the two largest categories in the table (mode 1 and 2) both have a telephone reinterview. The difference in *gdr* is small but not negligible, the cases involving personal interviews for the original interview have *gdr*'s about 10 percent greater than those conducted entirely by telephone. Mode 3 and 4 involve personal reinterviews and have much higher *gdr* than mode 1 and 2. As it is probably true that households for which the reinterview is conducted in person differ in many respects from the other households, no firm conclusion can be reached about the relative consistency of reports obtained by personal and telephone interviews. Nevertheless the results in table 8.9 suggest that, in terms of simple response variance, telephone interviewing is more reliable than personal interviewing -- the contrast between mode 1 and 3 indicates this.

The data in table 8.9 show one possibly important difference between the original interviews and the reinterviews. About one third of all CPS interviews are conducted face-to-face; fewer than 10 percent of the reinterviews involve personal visits. It might be useful to know more about the factors which determine the choice of mode of interview.

The three factors considered in this section have all been shown to be associated with differences in reliability of reporting. All of them are external to the role of the respondent. In section 8.4.2 a particular and important aspect of the respondent's role is examined -- the self/proxy status of the respondent -- using some of the results from this section.

8.4.2 Proxy- versus self-reports

In the CPS the interviewer is instructed to find an eligible respondent to provide information about the household; this individual then responds to the labour force questions for himself/herself and also for all other household members if they are not present. A question of general interest to survey researchers, and of specific interest to CPS researchers, is whether survey data suffer in quality to the extent that some sampled individuals do not respond for themselves (i.e., to the extent that proxy-response is accepted in lieu of self-response). Moore (1985) provides an interesting review of the literature in this area.

Because the CPS and its reinterview program do not control the identity of the respondent, and in particular do not control the self/proxy assignment, it is not possible to distinguish response status effects from effects due to differences in the samples in the two categories. It is nevertheless of interest to see to what extent the observed self and proxy groups differ in terms of simple response variance. From the respondent combinations available (see figure 8.3) the contrast between self/self *(P_0)* and proxy/same proxy *(P_4)* is the most appropriate. Table 8.10 shows the results for six labour force variables; in addition to the values of the *gdr*, the ratio of *gdr* for proxy-reports to self-reports is also given.

Table 8.10: Proxy vs. self-reports

Variable	P_4	P_0	Ratio = $\frac{gdr\ (P_4)}{gdr\ (P_0)}$
CLF	4.73	3.58	1.323
Employed	3.39	2.84	1.193
Unemployed	3.10	2.27	1.368
Non-Agric.	3.46	2.72	1.271
Full-time	3.75	3.63	1.031
Part-time	4.43	4.05	1.094

In general this table supports the conventional wisdom that proxy-reports are inferior to self-reports. The simple response variance for proxy-reports is of the order of 20-40 percent larger for proxy-reports for four of the variables; for "full-time" and "part-time" the difference

is slight. The comparison here is between individuals for whom the data were collected from themselves and those for whom the data were obtained from others but with the same respondent on the two occasions.

The results in section 8.4.1 indicated that there are important differences between the response variances for different types of respondents. In order to investigate whether this helps to explain the findings in table 8.10, the same ratio ($gdr(P_4)/gdr(P_0)$) was calculated separately for each of the subclasses considered in tables 8.7 and 8.8 for three key variables -- presence in civilian labour force (CLF), employed and unemployed. The results are shown in table 8.11. Each row of the table shows the values of the ratio for the three variables for a particular subclass; the overall ratio from table 8.10 is given for comparison in the final row.

Table 8.11: Proxy vs. self-reports for selected subclasses of subjects ratios of $\frac{gdr\,(P_4)}{gdr\,(P_0)}$

	Variable		
Subclass	CLF	Employed	Unemployed
Kin 1	0.51	0.66	0.43
Kin 2	0.68	1.11	0.45
Kin 3	1.08	1.09	0.97
Kin 4	0.99	1.25	0.76
Age 2	1.49	1.73	1.18
Age 3	0.98	0.93	0.92
Age 4	0.67	0.85	0.64
Age 5	0.47	0.38	0.61
Age 6	0.87	0.87	0.81
Age 7	0.91	1.02	0.51
All	1.32	1.19	1.37

These results are surprising. The values for the whole sample are very high compared to the values for each of the subclasses. In particular the majority of the values for the subclasses are less than 1.0, suggesting that for these cells the simple response variance for proxy-

reporting is less than that for self-reporting. A particular case may illustrate the problem.

Consider the variable "unemployed" and the subclasses kin 1, kin 2, kin 3 and kin 4; these subclasses together comprise virtually the whole sample. The ratio of the proxy/self simple response variances is less than 1.0 in every case, but the overall ratio is 1.37. The explanation of this phenomenon can be found by looking at the breakdown of the data given in figures 8.5A and 8.5B.

Figures 8.5A and 8.5B about here

Each of the pairs of values of *gdr* in figure 8.5A represents a category of subject. In all four sub-categories the *gdr* is larger for self-reports (P_0) than for proxy-reports (P_4); the overall values on the right represent all cases and there the *gdr* is larger for proxy-reports. The frequency distribution in figure 8.5B shows how this happens.

The distribution of individuals across subcategories is very different for self-reports from that for proxy-reports. For self-reports almost all the cases belong to kin 1 and kin 2 (heads of household) and kin 3 (wives of heads of household). For proxy-reports about 40 percent of the cases belong to the category kin 4 (other relatives). The kin 4 group has much higher *gdr* for both self and proxy-reports than any other category, and dominates the *gdr* for the proxy-report category. This is what leads to the erroneous impression, given by the overall comparison, that proxy-reports are less reliable than self-reports. The apparent contradiction between the marginal (overall) contrast and the detailed subclass contrasts is an example of Simpson's Paradox (Simpson, 1951). The important point to bear in mind is that the subclass contrasts represent a more controlled comparison of self-reports and proxy-reports. The apparent superiority of self-reports suggested by table 8.10 is merely an artifact of the relationship between the frequency distribution of response by type of subject and the different levels of reliability obtained for different types of subject.

The conclusion that proxy-reports are superior to self-reports is a little disturbing, as it contradicts the conventional wisdom that the best information comes from self-reports (see,

for instance, Sudman and Bradburn (1974)). It is not suggested that simple response variance is a measure that describes all, or even the most important, aspects of quality. It is possible that simple response variance may be lower but that biases, for instance, may be greater. There is no evidence in the data here to suggest that this is the case, but it is certainly a possibility that must be borne in mind. Mathiowetz and Groves (1983), however, in a study of telephone interviewing for the Health Interview Survey found "an overall tendency toward higher (better) proxy reports [which] runs directly counter to previous beliefs about self vs. proxy reports" (p. 96). The Mathiowetz and Groves results relate to completeness of reporting rather than variability in reports, and higher reporting is assumed to represent more complete, and therefore better, reporting. More recent research also supports the view that for factual data there is not necessarily a clear advantage to having self-reports.

A fundamental problem with the present results arises from the non-experimental nature of the selection of self and proxy respondents, and this suggests two other possible explanations. First, the respondents are self-selected and there may be biases in this process which invalidate the comparison. This would cause particular difficulty if the respondent characteristics that are the subject of inquiry are related to the factors which determine the likelihood of self-response (Moore 1985). There is some evidence to support this in the CPS data. The probability of being found at home, and therefore of being chosen as a respondent, is clearly related to labour force status and perhaps also to ambiguity of labour force status. This might apply particularly to kin 1, kin 2 and kin 4 in the tables above. Second, the interviewers also have some control over the choice of respondent. It is possible, though not testable, that judicious selection among eligible respondents may exaggerate the quality of data provided by proxy respondents.

There are two further possibilities which merit investigation. First, as demonstrated in section 8.3.1, both of the sets of estimates used in tables 8.10 and 8.11 are subject to recall effect. The model (table 8.3) specifies the same between-trial correlation for self and proxy reports. This may not be correct and a difference in between-trial correlation could explain part of the difference in observed simple response variance. Second, and this is related to the selection bias issues above, the model may still be misspecified and the explicit introduction of further

factors in the model might modify the results.

Notwithstanding the qualifications expressed above, the evidence presented in this section is sufficiently striking to warrant further controlled investigation. The acceptability of proxy respondents is an important practical issue in survey execution. The differences observed from the non-experimental data examined here justify the design and implementation of experimental studies to confirm or refute the conclusions drawn.

8.5 Combined Estimation and Reporting Issues

This section considers jointly some of the estimation issues discussed in section 8.3 and the reporting issues discussed in section 8.4 to illuminate both the interview process and some underlying characteristics of the respondents and the variables. Tables 8.7 through 8.9 in section 8.4 show how the *gdr* varies according to identifiable characteristics of the respondents and of the interview. The non experimental (i.e., non randomized) selection of respondents invalidates many comparisons. The categories used in section 8.4 are very broad; the category kin 1, for instance, includes all heads of household who have relatives in the same household. In order to make the comparisons more meaningful it is desirable to control as many characteristics as possible. In this section some additional controls are introduced.

8.5.1 Differential Communication and Contamination

By contrasting the *gdr*'s for P_1 with those for P_2, tables 8.6A and 8.6B in section 8.3.2 provided some evidence about within-household communication between the original interview and the reinterview. The P_1 cases are those for which the original interview uses a proxy respondent and the reinterview obtains a self-report; P_2 comprises those cases with a self-report in the original interview and a proxy-report for the reinterview. The assumptions are that communication between the two respondents is more likely to have an impact on the response when the self-report occurs first, that a ratio P_1/P_2 greater than 1 indicates communication and that its magnitude is a measure of the degree of communication. Since communication may have an affect even in the case of P_1, this may be considered a lower

bound for the effect of communication.

In order to reduce the effect of uncontrolled variables in the comparison, the data in table 8.12A are confined to husbands who are heads of households and their wives. There is some evidence of communication in the case of husbands; the values of the ratio P_1/P_2 are generally greater than 1, and substantially so for the variables "employed" and "unemployed". In the case of wives, there is hardly any evidence of communication at all; all the values of P_1/P_2 are closed to 1.0.

Table 8.12A: Differential communication for husbands and wives [effect measured by ratios of *gdr* (P_1) to *gdr* (P_2) for unreconciled data]

Variable	Married male head of household	Wife of head of household
CLF	1.16	1.15
Employed	1.53	1.11
Unemployed	1.97	0.99
Non-agric.	1.35	1.01
Full-time	1.01	1.13
Part-time	0.71	0.99

These data suggest that there is an asymmetry in communication in households. It would appear that wives are more likely to pay attention to (be influenced by) what their husbands say about being interviewed than husbands are to pay attention to (be influenced by) what their wives say.

Table 8.12B compares the effect of contamination for husbands and wives. These data refer to the 80 percent of the sample for which the interviewer has access to the original response. Here again P_1 is the group for whom the original response is a proxy response. It is hypothesized that the reinterviewer will give less weight to the original response in these cases than in the cases (P_2) where the original response is a self-response and the reinterview respondent is a proxy. If the hypothesis is correct the differential effect of contamination

should be greater for P_1 than for P_2; expressed another way, P_1/P_2 should be greater for contaminated than for unreconciled data.

Table 8.12B: Differential contamination for husbands and wives [effect measured by ratios of *gdr* (P_1) to *gdr* (P_2) for unreconciled data]

Variable	Married male head of household	Wife of head of household
CLF	1.95	0.97
Employed	2.02	1.01
Unemployed	1.73	0.86
Non-agric.	1.92	0.95
Full-time	1.63	1.01
Part-time	1.70	1.06

The evidence for husbands is strikingly different from that for wives. Where the husband is the subject and gives the original response the estimate of SRV (based on the difference between the two responses) is considerably lower than that for the case when the proxy-report comes first. This indicates that the response recorded for the reinterview is close to the original response when the original (self) response comes from the husband. The difference between the ratios P_1/P_2 in tables 8.12A and 8.12B indicates that part of this difference is due not to communication (table 8.12A) but to additional contamination.

The data for wives as subjects show a very different pattern. The ratios in table 8.12B indicate that the identity of the original respondent has no effect on the estimated SRV. The ratios P_1/P_2 are all very close to 1.0. This suggests that the reinterviewer does not allow an original response which is a self-report from the wife to have any more weight than an original (proxy) response from the husband.

It is possible to modify the model (table 8.3) in section 8.3 to include a term for differential communication and a term for differential contamination. The equation for the estimated SRV for P_2 will now include correlation coefficients analogous to the between trial correlation coefficient for recall which appears in table 8.3 for the expressions for P_0 and P_4. The

expressions for P_1 and P_2 therefore become:

For unreconciled data:

Category	Code	Estimate	Expected value
proxy/self	P_1	$s^2_{\varepsilon_1}$	$\frac{\left[\sigma^2_{\varepsilon(p)} + \sigma^2_{\varepsilon(s)}\right]}{2}$
self/proxy	P_2	$s^2_{\varepsilon_2}$	$\frac{\left[\sigma^2_{\varepsilon(s)} + \sigma^2_{\varepsilon(p)}\right]}{2}(1-\rho_c)$

(8.7)

For contaminated data:

Category	Code	Estimate	Expected value
proxy/self	P_1	$s^2_{\varepsilon_1}$	$k.\frac{\left[\sigma^2_{\varepsilon(p)} + \sigma^2_{\varepsilon(s)}\right]}{2}$
self/proxy	P_2	$s^2_{\varepsilon_2}$	$k.\frac{\left[\sigma^2_{\varepsilon(s)} + \sigma^2_{\varepsilon(p)}\right]}{2}(1-\rho_{cc})$

(8.8)

In equations (8.7) and (8.8), k is the general effect of contamination; ρ_c is the correlation representing the additional effect of communication on P_2 cases; ρ_{cc} is the combined effect of contamination and communication in P_2 cases. Tables 13A and 13B give estimates of ρ_c and ρ_{cc} for the husbands and wives data in tables 8.12A and 8.12B.

Table 8.13A: Estimates of communication effects ($\hat{\rho}_c$) for husbands and wives

Variable	Husbands + Wives	Wives + Husbands
CLF	0.14	0.13
Employed	0.35	0.10
Unemployed	0.49	-0.01
Non-agric.	0.26	0.01
Full-time	0.01	0.12
Part-time	-0.41	-0.01

For unreconciled data

$$\hat{\rho}_c = \frac{P_1 - P_2}{P_1}$$

Table 8.13B: Estimates of differential communication/contamination effects for husbands and wives

Variable	Husbands + Wives	Wives + Husbands
CLF	0.49	-0.03
Employed	0.50	0.01
Unemployed	0.42	-0.16
Non-agric.	0.48	-0.05
Full-time	0.39	0.01
Part-time	0.41	0.06

For contaminated data

$$\hat{\rho}_{cc} = \frac{P_1 - P_2}{P_2}$$

The conclusion to be drawn from the analysis in this section is two-fold. First, there is some quantifiable evidence that communication between respondents in the household is causing downward bias in the estimate of SRV, and that this communication is not equally effective in all cases. Second, the results indicate that not only do the contaminated reinterviews produce deflated estimates of SRV (as is well known) but that the extent of deflation varies considerably for different categories of respondent and subject. This further evidence about the effect casts doubt on the usefulness of the contaminated reinterviews.

8.5.2 The Measurement of Labor Force Status

The discussion so far has assumed implicitly that the variables being measured are equally relevant and clearly defined for all classes of respondents. This is unlikely to be the case.

Consider the hypothetical contrast between "a head of household employed full-time and salaried, and his daughter, a part-time student who spent a few hours last week (or was it the week before?) trying to sell cosmetics on commission without yet having collected any money for the effort" (Bailar and Rothwell 1984). The contrast here is between virtually error-free and clearly error-prone situations. Overall in the population the classification of employment status varies from robust and stable to weak and uncertain.

Table 8.14 presents the estimates of the *gdr* for the variable "unemployed" for selected subclasses of respondents and subjects. The first column of the table gives the GDR for reports on each class of subject by others, where the same proxy reported in the original interview and the reinterview. Thus, 1.26 is the estimated *gdr* for heads of household when both reports were by the same proxy-respondent. The second column of the table gives the estimated *gdr* for self-reports on both occasions. The third column gives the estimated *gdr* for reports by each class of respondent on others. Thus 4.25 is the estimated *gdr* for reports by heads of households on others. In each column both reports are by the same respondent.

Table 8.14: *gdr* for "unemployed" for selected subclasses of respondents and subjects: self-reports and proxy-reports

Subject	Proxy = 4	Self	Proxy = 4	Respondent
Kin = 1	1.26	2.94	4.25	Kin = 1
Kin = 3	1.67	1.72	2.56	Kin = 3
Kin = 4	5.67	7.48	3.90	Kin = 4
Age = 2	5.91	5.02	4.33	Age = 2
Age = 3	2.80	3.04	2.30	Age = 3
Age = 4	1.73	2.72	3.76	Age = 4
Age = 5	1.39	2.26	3.81	Age = 5
Age = 6	1.04	1.28	2.81	Age = 6
Age = 7	0.28	0.54	1.33	Age = 7

The contrast between column 1 and column 2 is that already discussed in section 4.2 -- the estimated *gdr* for self-reports is higher than that for proxy-reports for all but one of the classes. The contrast between columns 2 and 3 is the main concern here. In most cases the

estimated *gdr* in column 2 is less than that in column 3. This is what would be expected. Reports by an individual on himself/herself *should* be more reliable than an individual's reports on other people. Remember that the *respondents* in column 2 are the *same people* as the respondents in column 3, the *subjects* in columns 1 and 2 are merely the *same categories* of people; thus the contrast between columns 2 and 3 is much better controlled.

There are, however, three exceptions to the expected relationship between columns 2 and 3. These are the kin 4, age 2 and age 3 groups. For these, which comprise the younger, other relatives in household groups, the estimated *gdr* is higher when they report on themselves than when they report on others. These groups also have the highest *gdr*'s in column 1, when others report on them. This suggests that the problem with this group lies not in their capacity to report but in the clarity or robustness of the information that is being reported about them. This has two implications: first, there is a need for work on the problem of measurement or definition, at least for these individuals; and second, the unreliability of the response for these individuals is high regardless of the source from which the information is obtained.

The data permit one further controlled comparison of this kind. Husband and wife pairs can be identified. Tables 8.12 and 8.13 considered the proxy = 1 and proxy = 2 cases. Table 8.15 presents the values of *gdr* for proxy = 0 and proxy = 4. The contrasts in this table are of two kinds.

Table 8.15: Husbands and wives: proxy vs. self reports: values of *gdr*

Variable	Wife self reports (P_0)	Husband wife reports (P_4)	Husband self reports (P_0)	Wife husband reports (P_4)
CLF	3.89	1.64	2.58	4.09
Employed	3.07	1.86	3.02	3.27
Unemployed	1.72	1.19	2.05	1.73
Non-agric.	2.80	2.13	2.67	3.27
Full-time	3.65	3.21	4.18	3.91
Part-time	4.34	2.42	3.47	4.27

The contrasts between columns 1 and 4 and between columns 3 and 2 compare the same class of subject when the respondent is self (columns 1 and 3) and proxy (columns 4 and 2). These contrasts simply control the *type* of subject -- the individuals concerned are not the same for 1 and 4 or for 3 and 2, nor is the choice of respondent random.

In columns 1 and 4, wives are the subjects and the estimated *gdr*'s for self-reports are generally slightly lower than those for reports on wives by husbands. In comparing columns 3 and 2, where the subjects are husbands, the estimated *gdr*'s for self-reports are all higher than those for reports on husbands by wives. Although these two contrasts together suggest that wives are better respondents than husbands the observed differences may be due to the fact that those husbands who report at all (i.e., who become respondents) are likely to differ from those who do not. Those who are respondents are more likely to be at home during the day and may therefore be in a different type of employment or in a more ambiguous situation with regard to labour force status. For wives on the other hand, being at home may be less strongly related to ambiguity of labour force status.

The contrasts between columns 1 and 2 and between columns 3 and 4 are of a different kind from those discussed above. They compare the variability of responses by *the same respondents* for themselves and for their spouses. Thus the contrast between columns 3 and 4 compares self-reports by husbands with reports by those husbands on their wives. The estimated *gdr*'s are not very different although for "being in labour force" (CLF) and "part-time" the husbands' self-reports are better. When wives are the respondents, the pattern is different. For all the variables considered, wives' self-reports (column 1) are more variable than their reports on their husbands (column 2).

This latter result is similar to that observed in table 8.14 for the "other relatives" (kin 4) group. It also confirms the undesirability of assuming either that the pattern of variability is consistent across groups or that the groups are comparable simply because they have certain characteristics in common. It is incorrect to deduce from table 8.15 that wives know more about their husbands than their husbands know about them -- or even that they report what

they know more consistently. The results in this section indicate that there are interactions between the identifiable (demographic) characteristics of the respondents and the subjects, and that propensity to respond is related both to labour force characteristics and to the robustness and stability of labour force status.

8.6 Conclusions (and Recommendations)

The analyses in this chapter are of four kinds. The first set is concerned with the estimation of the simple response variance (SRV), and makes use of the design characteristics of the reinterview program to provide a modified estimator. The second set investigates the effects of communication and contamination on the responses in the reinterviews and uses these factors to shed light on the reinterview process. The third set examines the impact of the characteristics of the individuals and of the mode of reinterview on the SRV. The final set considers the implications of the earlier results for the design and implementation of the CPS itself. The eight recommendations in figure 8.6 are based on the results of the analyses.

The analysis in section 8.3.1 provides convincing evidence of the presence of a substantial between-interview correlation for the response deviations. The lower set of estimates given by equation (8.4) are conservative and estimates of SRV should be modified to take this between-interview correlation into account (recommendation 1.1). The evidence in section 8.5.1 about the correlation introduced by communication within households suggests that further modification may also be necessary. It would also be desirable to use a form of the estimator which is appropriate for cases selected with unequal probabilities.

Figure 8.6 Recommendations

1.1	Modify the estimator of SRV to include the effect of the between-interview correlations, ρ.
1.2	Reduce the proportion of the reinterview sample with contaminated responses.
1.3	Develop a model which includes both estimation characteristics (as covariates) and respondent/subject characteristics (as predictors).
2.1	Monitor SRV for different respondent groups.
2.2	Monitor the structure of the respondent set.
3.1	Carry out field experiments to test hypotheses about - choice of respondents - proxy vs. self - modes of interview
3.2	Carry out question design and implementation studies, particularly for problem categories of individuals.
4.	Evaluate impact of changes

The comparison of the contaminated (80 percent) and unreconciled (20 percent) data in section 8.3.2 demonstrates conclusively that the sample does not adequately represent the actual inconsistencies in the responses. Table 8.6 and tables 12B and 13B show that the effect of contamination is not uniform for different groups of respondents and subjects. The degree to which the reinterviewer is differentially influenced by the identity of the original respondent casts considerable doubt on the value of the contaminated interviews. Although these reinterviews are used to detect fabrication of data, and may have some value for interviewer training, the procedures should be examined to see whether the problems of execution can be overcome. In Computer Assisted Telephone Interviewing (CATI) and Computer Assisted Personal Interviewing (CAPI) it should be possible to provide access to the original responses only after the reinterview has been completed, and to prevent any subsequent changes in the reinterview responses themselves. Removing the contamination effect would greatly increase the data base for analysing response variance (recommendation 1.2).

The analyses in this paper deal separately with estimation issues and reporting issues. The next stage of analysis should involve the construction and estimation of a model which combines these factors (recommendation 1.3).

It is clear from the results in sections 8.4 and 8.5 that certain identifiable subclasses of respondents have different levels of SRV. The analysis of the reinterview program should include routine monitoring both of the structure of the respondent set (the proportion of respondents of different kinds) and of the SRV for these subclasses. It is possible that changes in data quality may occur which apply only to particular subclasses, or that changes may occur which are due simple to shifts in the balance between good and bad respondents (recommendations 2.1 and 2.2).

In order to reach valid conclusions about the relationships between type of respondent and data quality, it is necessary to control other factors which might be confounded with the factor being examined. Thus, in comparing self-reports and proxy reports, for instance, the contrast will be invalidated if other characteristics of the respondents besides their self/proxy status vary for the two groups. In order to overcome this problem the allocation to self or proxy status must be randomized. The same argument holds for the other contrasts in section 8.4. The differences observed in the analyses in this paper should be tested by carrying out field experiments to examine in particular the mode of interview, proxy vs. self reports, and choice of respondent in the household (recommendation 3.1).

The results in section 8.5.2 demonstrate that for some subclasses of individuals there is evidence that the ambiguity or instability of their labour force status is such that the questions and procedures being used in the CPS are not adequate to ascertain the information reliably. This is particularly striking for the group of "other relatives" (kin 4) who report more reliability for others than they do for themselves. It is therefore important that further work be done to identify these subclasses and to design questions and procedures that will permit better determination of their labour force status (recommendation 3.2).

It is essential that for each of the recommendations above, and for any other changes arising

from them, the impact of implementing the changes be monitored and that appropriate adjustments be introduced where necessary (recommendation 4). It is only by continued development and evaluation that real improvements in data quality can be achieved.

Chapter 9 - MODELLING OF SRV FOR BINARY DATA

9.1 Alternative Models for Response Variance

The design and estimation issues described in section 8.2 of chapter 8 are almost inextricably intertwined. The purpose of this chapter is to provide a framework within which they can be disentangled.

The basic characteristic in the examination of response errors is the response deviation d_{jt}. In the context of the response variance the dependent variable is ε_{jt}, the variable response error. Our objective is to explain ε_{jt} in terms of the characteristics of the individual j, or of the characteristics of the respondent (if the respondent is different), or of other aspects of the measurement process.

It is difficult to model ε_{jt} because under the model it is a random variable with an expected value of 0. Consequently we first considered using as the basis of the analysis the quantity (*equation*), the *SRV*, or the absolute value of the response error ε_{jt}. It is difficult, however, to model variation in either of these quantities satisfactorily. In the first place it is difficult to know whether an additive or a multiplicative model is the more appropriate; as the ideal value of $|\varepsilon_{jt}|$ would be 0, if we want to model departures from zero an additive model would seem to be necessary, whereas a multiplicative model is intuitively more appealing.

The problems are exacerbated when we consider the case of binary variables. The response deviation d_{jt} can take only three possible values, -1, 0 and +1, and the response process can be defined by the probability of a false positive (φ) and the probability of a false negative (θ). The data set can also be described completely as in figure 8.1 in chapter 8.

Consider a large table cross-classifying the possible explanatory variables and representing the sample in a multi-way contingency table. Such a table has in each cell the number of cases (frequency) with that combination of characteristics. Our interest, however, is not in the frequency in the cell but in the response variability demonstrated by the elements in the cell.

In other words we have for every cell the 2×2 table representing the responses on the two trials. What we would like to do is to model the variation among the many 2×2 tables in terms of explanatory variables where these are the defining categories for the cell.

In the final section of this chapter we discuss various possible ways of reparameterizing the 2×2 table. Here we consider a particular way of summarizing the information contained in the table. We already have a summary measure available to describe a 2×2 table. This is the *gdr* [=(*b*+*c*)/*n*], which is of course equal to twice the estimate of the *SRV*. It is the proportion of disagreements between $t = 1$ and $t = 2$ and must take a value between 0 and 1. The measure itself therefore suffers from the same disadvantages as the *SRV* as the dependent quantity in modeling the response variance.

Writing the *gdr* as

$$g = (b+ c)/n$$

the odds of disagreement between the observations for $t = 1$ and $t = 2$ can be written as

$$g/(1-g).$$

These odds vary from cell to cell. What we are trying to do is to explain this variation in terms of the identifying characteristics of the cells (i.e. the characteristics of the individuals in the cells or the measurement process represented by the combination of factors denoted by the cells). We postulate multiplicative effects on the odds. A simple example would be that the effect of being, say, younger multiplies the odds of disagreement between the first and second measurements by some factor.

It may be worth pointing out at this stage that there is no restriction on the factors which can be included in the model. They may include both the factors affecting the response variance itself (the left hand side of figure 8.2 in chapter 8) and the factors affecting its estimation (the right hand side of figure 8.2). Since the model is multiplicative in the odds it is easier to express it as an additive model in the logarithms of the odds. Thus for a model with eight explanatory variables *A, B, C, ..., H* the model can be written as

$$\log_e\left[\frac{g_{ijk\ldots p}}{1-g_{ijk\ldots p}}\right] = \beta^1 + \beta_i^A + \beta_j^B + \beta_k^C + \ldots + \beta_p^H + \beta_{ij}^{AB} + \ldots etc \qquad (9.3)$$

where $\hat{\beta}_i$ represents the additive effect on the log odds of belonging to category *i* of variable *A*, etc.

9.2 The Data and the Estimation Procedure

In the form in which this model is set up, the data set being analyzed is the multiway table where the element within each cell is the ratio $g/(1-g)$. The model is an asymmetric model with a single dependent variable; it is in fact a logistic regression - a logit model with binomial error. There is a slight violation of the assumptions involved here as the data come from a process in which the odds of disagreement arise from two different sources - the probability of a false positive and the probability of a false negative.

β_i^A

As the data set consists of a table it is necessary to construct the table as the first stage in the analysis. The basic unit is the cell, and the quantity $g/(1-g)$ must be calculated for each cell. The analysis will then fit the specified model to the data and calculate the degree to which the model fits the data. This raises three issues. The first is the problem of empty cells in the table; the second is the evaluation of the model; the third is the interpretation of the results in terms of the original (raw) data.

The analysis was carried out using the GLIM program. The raw data set was very large (126,122 cases) and it was not convenient within GLIM to process the raw data. Thus each cross-tabulation was first produced outside GLIM and imported into the program. This meant that the set of variables for each analysis had to be determined in advance and the appropriate cross-tabulation produced.

For the examples presented in this chapter, the set of explanatory variables considered was confined to the set used in chapter 8. This set, described in figure 9.1, provides an adequate range of variables and also permits a comparison of the results and their interpretation with

the previous work.

Figure 9.1: Explanatory variables used in these analyses.

Contamination (CONT)

1 Yes

2 No

Response status (PROX)

see figure 8.3 in chapter 8

Relationship to head of household (KIN)

1 Head of household

2 Spouse of head

3 Other relative

The same categorization for the *respondent* at the first and second interview is denoted by R1KIN and R2KIN

Mode of interview and reinterview (MODE)

1 Telephone/telephone

2 Face-to-face/telephone

3 Face-to-face/face-to-face

4 Telephone/face-to-face

5 Mixed in either interview

Age of subject (AGE)

1 24 years and under

2 25-34

3 35-44

4 44-54

5 55-64

6 65 years and over

Example 1

The first example is used simply to illustrate the application of the model and the interpretation of the parameters. The data set consisted of a cross-tabulation of the data by CONT, PROX, KIN, R1KIN, R2KIN and MODE and thus contained 2×5×5×5×5×5 = 6,250 cells.

Inspection of the list of explanatory variables makes it clear that some cell of the table contain

what are called structural zeros, i.e., by definition it is impossible that there should be any observations in these cells. For example if PROX = 1 (self-reports on both occasions), then KIN must be equal to both R1KIN and R2KIN and the other cells in the subtable defined by PROX = 1 must be empty. In the analysis these cells must be specified as structural zeros in order to produce appropriate estimates for the parameters. It is also possible that there will be zero cells simply because of sampling fluctuations; these are called sampling zeros and are not specified in advance - the program makes appropriate allowance for them in the standard analysis. It is of course desirable that there should be few cells with low expected frequencies.

As a first stage in carrying out the analysis we fit the null model to the data. This essentially corresponds to finding out how much variability there is in the data and provides a benchmark against which to compare the fit of other models. The model could be specified as

$$\log\left[\frac{g}{1-g}\right] = \beta^1 \tag{9.4}$$

Of the 6,250 cells in the cross-tabulation 1,075 were nonempty. Thus there are 1,074 degrees of freedom available in the table. This model simply fits a constant. The null model was fitted in the case of both the dependent variables, whether or not in the Civilian Labor Force (CLF) and whether or not unemployed (UNM).

The scaled deviance, which is a measure of the variation between the fitted and observed values in the cells of the table was 2,971 for CLF and 2,189 for UNM..

To illustrate the procedure we fit the main effects model

$$\log\left[\frac{g}{1-g}\right] = \beta^1 + \beta_2^{CONT} + \beta_2^{PROX} + \beta_3^{PROX} + \beta_4^{PROX} + \beta_5^{PROX} \tag{9.5}$$

In this model we allow the odds of disagreement between the original interviews and the reinterviews to be affected by the explanatory variables CONT and PROX. It is worth noting that the number of fitted terms for each of the explanatory variables is one less than the number of categories for that variable. Thus for CONT, which is binary, there is one term; for PROX, which has five categories, there are four terms. These two explanatory variables thus use up five degrees of freedom [(2-1)+(5-1)].

The scaled deviance for the model was 2,307 for CLF and 1,788 for UNM with 1,069 degrees of freedom. The change in deviance is therefore 664 for CLF and 401 for UNM and the change in degrees of freedom is 5 in each case. Although the deviance itself cannot be tested directly, the difference between the deviances for the two models can be tested using chi-square. The two explanatory variables clearly contribute significantly to the explanation of the variation in the odds of disagreement.

Having established that the explanatory variables warrant inclusion in the model, table 9.1 presents the parameter estimates and their estimated standard errors for the two dependent variables CLF and UNM. The sizes of the estimated standard errors relative to the parameter estimates reflect the conclusion above about the strongly significant effect of each of the explanatory variables.

Table 9.1 Estimates for the model incorporating the main effects of contamination and self/proxy response status

	Dependent Variable			
	CLF		UNM	
Parameter	Estimates	s.e.	Estimates	s.e.
β^1	-3.94	0.034	-4.48	0.043
β_2^{CONT}	0.70	0.037	0.73	0.046
β_2^{PROX}	0.57	0.058	0.65	0.072
β_3^{PROX}	0.36	0.063	0.36	0.081
β_4^{PROX}	0.20	0.042	0.30	0.053
β_5^{PROX}	1.08	0.059	0.96	0.078

The parameter estimates represent the deviation from the reference category for each variable. In these examples the reference category is the first category specified - for CONT it is the contaminated cases, for PROX it is the self/self category.

Taking as an example the parameter estimate β_2^{CONT}, the effect on the odds of disagreement of the reinterview not being contaminated is that the log odds are increased by 0.70 (CLF) or 0.73 (UNM). The odds are therefore multiplied by $e^{0.70}$ or $e^{0.73}$, i.e., 2.01 or 2.08. In other words the odds of disagreement between the first and second responses are approximately twice as high in the case of an uncontaminated reinterview as in the case of a contaminated one; put another way, the contamination depresses the estimate of response variability by about 50 percent.

9.3 The Effects of Many Variables

The characteristics of the respondents in a survey have been shown to be related to the quality of the responses obtained. For fertility data, for instance, O'Muircheartaigh (1984) showed that respondent age and education were related to the size of the simple response variance. In addition to these extra-role characteristics of the respondent, the self/proxy response status may be important, as may the relationship between the respondent and the subject.

For CPS data the identity of the respondent is included on the record for each subject (the subject being the individual about whom the survey variables are being reported). Although the respondent's *characteristics* are not included on the records for the subjects for whom he or she responds, it is possible to obtain the information from the respondent's own subject record. This was done for the full data set.

Five factors in all are considered here: two subject-related factors - the relationship of the individual to the head of household (KIN) and the age of the individual (AGE); two interview-related factors - mode of interview and reinterview (MODE) and whether the reinterview was contaminated (CONT); and one reporting factor - the self/proxy status of the interview and the reinterview.

In the analysis carried out in chapter 8 each of these factors was considered; the tables used were confined to uncontaminated self-reports. The tables were presented in terms of *gdr* for each of the categories of respondent. Table 9.2 below extracts results for the two dependent variables for KIN, AGE, and MODE. In this chapter *gdr* is expressed as a proportion rather than as a percentage (as in chapter 8), as this is the most appropriate formulation for the logistic models we will be using here.

Table 9.2 Gross difference rate for CLF and UNM by relationship to head of household (KIN), age, and mode of interview and reinterview (MODE)

	Dependent	
Explanatory	CLF	UNM
KIN		
Head of household	0.036	0.029
Wife of head of household	0.039	0.017
Other relative	0.082	0.075
AGE		
24 and under	0.058	0.050
25-34	0.030	0.030
35-44	0.042	0.027
45-54	0.037	0.023
55-64	0.029	0.013
65 and over	0.031	0.005
MODE		
Telephone/telephone	0.027	0.016
Face to face/telephone	0.031	0.017
Face to face/face to face	0.050	0.044
Telephone/face to face	0.056	0.034
Mixed modes in either interview	0.033	0.018

There is evidence from the table that consistency of reporting is higher for older respondents in the CPS. The results for UNM are those which show the clearest pattern. The estimated *gdr* decreases from a level of 5.02 per cent for the youngest age group (age 24 and under) to

a level of 0.54 per cent for the oldest group (65 and over) and there is a monotonic trend between these extremes. For the labor force status variables generally, the highest variability occurs for the younger respondents and the lowest for the older respondents.

In the case of relationship of the respondent to the head of household (KIN) the most striking aspect is the very high *gdr* found for the 'other relatives' group. The values for this group are generally about twice as large as for any other group and are worst for the two variables considered here - CLF and UNM.

In the case of MODE, the interest was in the comparison of face to face and telephone interviewing. Much of the interviewing in the CPS is done by telephone (from the interviewer's home) although the choice of mode is not randomized. In assessing the effect of mode, the cases must be classified by the combination of modes on the two occasions.

Since an even higher proportion of reinterviews is conducted by telephone, the two largest categories (modes 1 and 2) in the table both have a telephone reinterview. The difference in *gdr* is small but not negligible, the cases involving face to face interviews for the original interview have *gdr* about 10 percent greater than those conducted entirely by telephone. Modes 3 and 4 involve face to face reinterviews and have much higher *gdr* than modes 1 and 2. As it is probably true that households for which the reinterview is conducted in person differ in many ways from the other households, no firm conclusion can be reached about the relative consistency of reports obtained by face to face and telephone interviews. Nevertheless the results suggest that telephone interviewing is at least as reliable as face to face interviewing.

When separate logit models were fitted to the data, including each of the explanatory variables above one at a time, the results naturally confirmed the results above. This is so because the logit model simply reexpresses the relationship in the form of odds of disagreement rather than the proportion of disagreements. In terms of one-way or marginal analyses, therefore, the implications of the results will necessarily be the same.

The principal advantage of the log-linear formulation is that it permits the joint analysis of the effects of many variables. With the *gdr* as the measure of variability and using the cross-tabulation as the method of comparing effects, joint analysis by many variables quickly runs into difficulties in terms of cell sizes and the stability of the estimates. Once the model is specified as a logistic regression, however, there is no difficulty in estimating simultaneously the effects of many variables. For the dependent variable UNM we fitted a model which included six explanatory variables - AGE, KIN, CONT, PROX, R1KIN, and R2KIN. [The data were confined to three categories of KIN - (male) heads of household, wives of heads of household, and other relatives in the household, in order to simplify the interpretation of the results.]

$$\log\left[\frac{g}{1-g}\right] = \beta^1 + \beta^{AGE} + \beta^{KIN} + \beta^{CONT} + \beta^{PROX} + \beta^{R1KIN} + \beta^{R2KIN} \qquad (9.6)$$

Table 9.3 shows the estimated parameters for the model (9.6) and also for each of the explanatory variables fitted independently. Column 1 of the table shows the marginal effect of each explanatory variable; column 2 shows the estimated effect when all six explanatory variables are included jointly in the model - i.e., the effects of each variable taking into account (or controlling) the effects of the other variables. The model given here includes only main effects and does not allow for the presence of interactions between the explanatory variables.

Table 9.3 Estimated parameters for the logistic regression of odds of disagreement on the set of explanatory variables, fitted separately and jointly

Marginal effect			Joint effects	
Estimate	e^{β}	Parameter	e^{β}	Estimate

0.00	1.00	β_1^{AGE}	1.00	0.00
-0.71	0.49	β_2^{AGE}	0.67	-0.40
-1.07	0.34	β_3^{AGE}	0.51	-0.68
-1.28	0.28	β_4^{AGE}	0.41	-0.88
-1.60	0.20	β_5^{AGE}	0.30	-1.21
-2.96	0.05	β_6^{AGE}	0.07	-2.64
0.00	1.00	β_1^{KIN}	1.00	0.00
0.07	1.07	β_2^{KIN}	0.94	-0.06
1.19	3.29	β_3^{KIN}	1.90	0.64
0.00	1.00	β_1^{CONT}	1.00	0.00
0.75	2.10	β_2^{CONT}	2.08	0.74

In the case of CONT, the effect of contamination on the estimate of response variability, the presence of the other variables in the model has no appreciable effect on the estimate of the parameter. Both when it is considered alone and jointly with the other variables the impact of contamination is to halve the estimated response variability.

The second and third explanatory variables considered in table 9.3 are the age of the subject (AGE) and the relationship of the subject to the head of household (KIN). These variables are the most relevant variables of this kind available in the data set. Considering KIN in isolation we see that heads of household have the lowest response variability, wives of heads of household have almost the same, but other relatives in the household have a considerably higher response variability - their odds of discrepancy between the two interviews being more than three times higher than the other two groups. In the case of AGE there is a clear monotonic relationship between the age of the subject and the odds of discrepancy. The ratio of the odds for the youngest to the odds for the oldest is almost 20 to 1.

When the other predictors are taken into account, the pattern remains the same but the magnitude of the effects changes. Heads of household and their wives can still be seen to have comparable response variability and other relatives still have much higher response variability. The ratio of the odds of discrepancy for other relatives to the odds for heads of household and spouses is reduced to a little under two to one when age, etc. are taken into account. In the case of AGE the monotonic effect persists but the magnitude of the effects

is reduced. The ratio of the odds for the youngest to the oldest is now a little under 14 to 1.

The analysis described above carries us a good deal farther in terms of understanding the determinants of response variability than previous analyses. The main advantage is that it is possible to consider the joint effects of many variables rather than confining ourselves to the analysis of the correlates of response variability one at a time. In the case of the joint analysis of AGE and KIN it is clear that, whereas there is an association between the two variables, each also has an effect separate from the other.

The other important variable fitted in the model above is PROX, the variable which measures the self/proxy response status for each of the responses in the data set. As this variable raises a number of particular issues, it is treated in the next section.

9.4 The Effect of Proxy Reporting

Consider again the interview situation as described earlier in figure 8.2 in chapter 8. The data for some individuals (subjects) will be based on self-reports and for others will be based on reports by others (proxy reports). There is no prior basis for believing that the reliability of reporting will be the same for these two situations. Indeed, the conventional wisdom is that self-reports will be superior to proxy reports. A question of general interest to survey researchers, and of specific interest to CPS researchers, is whether survey data suffer in quality to the extent that some sampled individuals do not respond for themselves (i.e., to the extent that proxy responses are accepted in lieu of self-responses). Moore (1985) provides an interesting review of the literature in this area.

Whether a report is a self-report or a proxy report is a factor which may affect the quality of the report itself. In estimating the quality of the report, however, we need also take into account the extent to which the reinterview procedure departs substantially from the ideal of independent replication. Apart from procedural differences (including contamination, which affects the interviewer) there are two factors which may create a dependence between the response deviations in the first and second interviews. First, if the same individual responds

to the questions on the first and second interviews, there is a chance that he or she will tend to give the same response on the two occasions because he or she remembers the response given on the first occasion. This would introduce a positive correlation between the response deviation on the original interview (ε_{j1}) and that on the reinterview (ε_{j2}). Second, even if the respondents are different on the two interviews, communication among members of the household could lead similarly to a positive between-interview correlation for the response errors.

In the CPS any eligible respondent can report for the whole household in either interview; thus some observations are self-reports and some are proxy reports. In terms of the interview-reinterview data being used in the estimation procedure, five categories of self/proxy combinations may usefully be identified - these are given in figure 8.3 of chapter 8.

In terms of the U.S. Bureau of the Census model of response variance, the impact of either memory or communication can be represented by a correlation coefficient ρ. A number of attempts have been made to estimate the value of ρ; these are described in chapter 8. Using the data considered in this chapter, presented a model for the estimation of ρ is also presented in chapter 8. The model provided two different quantitative estimates for ρ. The model postulated that the simple response variance was different for self-reports and for proxy reports and that there was a between-interview correlation only for the reinterviews in which the respondent was the same as the respondent in the original interview. The correlation size was assumed to be the same for self- and proxy reports.

Two conclusions emerged from that analysis. First, the evidence is strong that there is a non-negligible between-trial correlation and that the usual estimate significantly underestimates the true *SRV*. The lower estimates assume that the mixed self/proxy cases are free from correlation. To the extent that there is communication within the household between the two respondents these estimates will be too low, thus it may be reasonable to assume that adjusting the estimates of *SRV* will underestimate *SRV*. Second, it is clear from the divergence between the two estimates of ρ that the simple model is inadequate. No

parsimonious model will ever represent the data perfectly, but it is clear that this model ignores some important terms.

One possibility is to allow the between-trial correlation to differ for self-reports and proxy reports. This unfortunately does not help, as the data will not produce valid estimates of the two correlations jointly (i.e., estimates with $|\rho \leq 1|$). This suggests, as does an inspection of the data, that other characteristics of the individuals besides their self/proxy status must be introduced into the model.

The analysis in section 8.4 of chapter 8 shows that two important factors in determining the response variability of an observation are the age of the subject and the relationship of the subject to the head of household. Both these variables are related to the likely labor force status of the subject and also to the stability of that status.

In this section we reconsider the findings from chapter 8 using the model proposed in this chapter. The analysis is confined to the variable UNM.

Table 9.4 gives the parameter estimates derived from the model

$$\log\left[\frac{g}{1-g}\right] = \beta^1 + \beta_1^{PROX} + \beta_2^{PROX} + \beta_4^{PROX} + \beta_5^{PROX} \qquad (9.7)$$

and the corresponding values of *gdr* obtained from the 1986 analysis (see chapter 8).

Table 9.4 **Values of *gdr* for the five self/proxy combinations, together with the logistic parameter estimates for the marginal effect of self/proxy response status, for the variable UNM**

Category	*gdr*	Estimate	e^{β}
self/self (P_0)	0.023	0.00	1.00
proxy/self (P_1)	0.041	0.52	1.68
self/proxy (P_2)	0.031	0.25	1.28
proxy/same proxy (P_4)	0.031	0.22	1.25
proxy/different proxy (P_5)	0.064	0.88	2.41

The results are fairly easy to interpret. The pattern of variation exhibited by the results is informative (and is similar for the five labor force status variables not presented here). In terms of *gdr* the lowest values of response variability are found in the P_0 (self/self) category, and these are uniformly lower than the values for the P_4 (proxy/same proxy) category. The two mixed categories P_1 (proxy/self) and P_2 (self/proxy) have similar levels of *gdr* for most variables, with a tendency for P_1 to be greater than P_2. The values of *gdr* for P_5 are very large relative to the other four categories.

In terms of the logistic model there is a parallel interpretation of the parameter estimates. The smallest value of beta occurs for P_0, and P_1 and P_2 lie above P_0 and P_4. When expressed in this way, it is still clear that the response variability for P_5 is considerably larger than for the other categories.

This univariate analysis fails to take into account the other characteristics of the subjects and the respondents which may affect the response variability of the observations. From section 8.4 we know that, in particular, the age and relationship to the head of the household of the subject are related to the degree of response variability. By fitting the model in equation (9.6) we can examine the way in which including these variables in the model (i.e. controlling for the effect of these variables) changes the interpretation of the effect of self/proxy response status. Table 9.5 gives the parameter estimates obtained when the six explanatory variables are fitted simultaneously. For comparison, the parameter estimates from fitting the one-

variable model are given in the first column of the table.

Table 9.5 Estimates of effect of self/proxy response status for UNM

	Marginal effect		Effect in (9.6)	
Category	Estimate	e^{β}	Estimate	e^{β}
P_1 (self/self)	0.00	1.00	0.00	1.00
P_2 (proxy/self)	0.52	1.68	0.32	1.38
P_3 (self/proxy)	0.25	1.28	0.07	1.07
P_4 (proxy/same proxy)	0.22	1.25	-0.39	0.68
P_5 (proxy/different proxy)	0.88	2.41	-0.20	0.82

The first point to note is that the relative response variability for the five self/proxy categories has been completely changed. The lowest variability is now to be found for P_4 (proxy/same proxy), and the second lowest for P_5 (proxy/different proxy). The three categories which include one or more self-reports have the highest response variability.

This reversal of the apparent effect of proxy reporting comes about because the analysis in table 9.5 of the contrast between response status categories is controlled for some other characteristics of the data and these other characteristics are related to the level of response reliability. In particular the great majority of self-reports are from heads of household and wives of heads of household; these are types of respondents with relatively low response variability. Among proxy reports, on the other hand, about 40 percent of the cases belong to the "other relatives" category; this is a category with relatively high response variability. This imbalance in the distribution of self/proxy classes produces an artificially high estimate of response variability for the cases with proxy reports in one or other interview. The impact is particularly striking for the P_4 and P_5 groups, in which both interviews have proxy respondents.

This means that the marginal effect of proxy reporting (the effect estimated by a simple comparison of the response variabilities of the different classes of self/proxy status without taking the other characteristics of the cases into account) creates an entirely false picture of

the true situation. This is an example of Simpson's paradox (for discussion see chapter 8). The apparent superiority of self-reports suggested by table 9.4 is merely an artifact of the relationship between the frequency distribution of self/proxy response status by type of subject and the different levels of reliability obtained for different types of subject.

In order to disentangle the influences at work here it is useful to reexpress what is happening in different terms. The first factor we would like to evaluate is the contrast between self and proxy reporting; thus we would like to estimate a self and a proxy effect. A factor which contaminates our estimate of this effect is memory (when the two occasions have the same respondent) and this memory effect may be different for self reports and for proxy reports. It is not unreasonable to postulate that memories of self reports will be stronger than memories of proxy reports.

The second factor is communication within the household. Communication is clearly possible in all situations. Of the three nonmemory situations, communication is least likely to be a factor in the case of P_5, where there are two different proxies; in any case the effect of communication is confounded in the estimation process. Of the two remaining situations we distinguish here between the case where the first report is a proxy report and the case where the first report is a self report. Where the first report is a self report, we argue that the impact of the communication may be to impress on the mind of the subsequent proxy respondent the answer(s) given by the self respondent so that there will be a positive between-interview correlation for the response deviations. Where the first report is given by a proxy respondent there is much less likely to be an effect of the communicated response on the subsequent self report.

To clarify the reasoning involved we express the five classes of self/proxy reports in different terms. Using notation analogous to that in the logistic model we will denote

effect of self-reporting by β_s
effect of memory (self) by $\beta_{m(s)}$
effect of proxy reporting by β_p

effect of memory (proxy) by $\beta_{m(p)}$

effect of communication by β_c.

We can now write the relationship between the five parameters β_1^{PROX} through β_5^{PROX} and the parameters above as

$$\beta_1^{PROX} = \beta_s + \beta_{m(s)}$$
$$\beta_2^{PROX} = (\beta_p + \beta_s)/2$$
$$\beta_3^{PROX} = (\beta_s + \beta_p)/2 + \beta_c$$
$$\beta_4^{PROX} = \beta_p + \beta_{m(p)}$$
$$\beta_5^{PROX} = \beta_p$$

where we expect the parameters $\beta_{m(s)}$, $\beta_{m(p)}$, and β_c to be negative.

Substituting the estimated values for β_1^{PROX} through β_5^{PROX} we can solve these equations for the five new parameters. We obtain the values

$$\beta_s = 0.84 \Rightarrow e^{\beta_s} = 2.32$$
$$\beta_p = -0.20 \Rightarrow e^{\beta_p} = 0.82$$

for self and proxy reporting.

The analysis confirms, as did the earlier analysis, that self reporting is less reliable (more variable) than proxy reporting. This is disturbing, as both the conventional wisdom and common sense suggest that the best information comes from self reports. A fundamental problem with the present results arises from the nonexperimental nature of the selection of the self and proxy respondents. Respondents are self-selected and there may be biases in this process which invalidate the comparison. This would cause particular difficulty if the respondent characteristics that are the subject of the inquiry were related to the factors which determine the likelihood of self response. There is some evidence in the CPS data to support this. The probability of being found at home, and therefore of being chosen as a respondent, is clearly related to labor force status and perhaps also to ambiguity of labor force status. Furthermore the interviewers also have some control over the choice of respondent. It is

possible, though not testable, that judicious selection among eligible respondents may exaggerate the quality of data provided by proxy respondents.

It is not suggested that response variability is a measure that describes all, or even the most important, aspects of data quality. It is possible that the response variability may be lower but that biases, for instance, may be greater. There is no evidence of this in the data here, but it is certainly a possibility.

There is an accumulation of evidence (see also chapter 8, section 8.4) that suggests that in many situations self reports are not as clearly superior to proxy reports as we have previously imagined. It seems counter-intuitive that proxy reports should in general be superior other than in situations where there is a strong social desirability or threat element in the questions. Notwithstanding the qualifications expressed above, the evidence is now sufficiently striking that further controlled investigation is warranted. One hypothesis we would put forward is that the design of the survey instrument and its administration is not motivating the self reporter sufficiently to take advantage of the superior information available to him/her.

For the memory and communication effects we obtain the values

$$\beta_{m(s)} = -0.84 \Rightarrow e^{\beta_{m(s)}} = 0.43$$
$$\beta_{m(p)} = -0.19 \Rightarrow e^{\beta_{m(p)}} = 0.82$$
$$\beta_c = -0.25 \Rightarrow e^{\beta_c} = 0.78$$

If transformed to the parametrization of the earlier model these estimates would give

$$\rho_{m(s)} = 0.57$$
$$\rho_{m(p)} = 0.17$$
$$\rho_c = 0.22.$$

The estimates of $\beta_{m(s)}$, $\beta_{m(p)}$, and β_c clarify considerably the impact of the factors on the right hand side of figure 8.2. First, they are intuitively acceptable (or plausible) values. The highest

effect is for self memory; we would have expected this to be the case as a self report is more salient to the individual than would be a proxy report. The proxy memory effect is considerably lower but is still substantial. The effect of communication is very similar to the proxy memory effect.

It may be instructive to give an interpretation of the parameters under each of the two models. The correlation coefficients $\rho_{m(s)}$ etc. express the memory and communication effects as a suppression of the simple response variance due to correlation between the response deviations on the two occasions. Thus the value of $\rho_{m(s)}$ of 0.57 implies that the correlation between the response deviations is such that when self reports are obtained on both occasions the estimated *SRV* underestimates the true *SRV* by 57 percent. When the two responses are from the same proxy respondent, the underestimation is of the order of 17 percent. Even when the two respondents are different, communication between the respondents on the two occasions can lead to a 22 percent underestimation of the *SRV*.

Under the log-linear model, the memory effect is seen as a reduction in the odds of finding a discrepancy between the responses on the two occasions. In a manner analogous to that in the *SRV* model, the fact that the questions are answered by the same individual on the two occasions means that the true odds of discrepancy will be underestimated by the proportion of discrepancies in the data. A value of $\beta_{m(s)}$ of -0.84, giving a value of 0.43 for e^{β}, implies that the expected value of the observed odds of discrepancy will be less than half the true odds of discrepancy. Similarly, the values of -0.19 for $\beta_{m(p)}$ (0.82 for e^{β}) and -0.25 for β_c (0.78 for e^{β}) quantify the suppression of discrepancies by proxy memory and communication, respectively.

Though it is possible to interpret the memory and communication effects under either of the two models, it is worth pointing out that the Census Bureau model would not permit the direct estimation of the parameters because of the need to take into account the joint effects of many variables in the estimation.

9.5 Conclusion

For more than thirty years the response variability in the Current Population Survey has been measured on the basis of the gross difference rate, *gdr*, presented by Hansen and his colleagues in 1961. Dealing as they were primarily with binary data the *gdr* provided a plausible and intuitive measure of the reliability of the responses. The *gdr* is the proportion of cases in which the reinterview response is different from the response in the original interview. It is also equal to twice the simple response variance (*SRV*) and therefore fits well into the usual survey model of response errors. Two particular issues are addressed in this chapter. The first is the problem of estimating the quality of the data given the shortcomings of the reinterviews as independent replicates of the original interviews. The second is the need to understand the determinants of data quality.

It happens that these issues are interrelated, and that in order to tackle them it is necessary to express the response process in a different way from that in which it is expressed in the Census Bureau model. The crucial difference is that instead of expressing response variability in terms of the proportion of discrepancies between the two interviews, it is expressed in terms of the odds of discrepancy. Though this may appear to be a slight difference it has very considerable implications for the analysis of data quality. In particular it makes it possible to construct a model in which both the factors affecting data quality and the factors affecting the measurement of data quality can be incorporated. We argue that this model is both more plausible and more useful than the standard model. It addresses the fundamental problems which arise in dealing with binary data in that the model is one specifically designed for such data and it makes it possible to incorporate into the analysis the structure of the data collection process.

The model provides an advance on earlier work in that it makes it possible to address all these issues simultaneously. It produces estimates of memory effects and communication effects, and distinguishes between memory for self reports and proxy reports. At the same time the analysis confirms the characteristics that are related to the magnitude of response variability, while controlling for the measurement process.

The analysis also provides some ideas for future work. It should be possible to

reparameterize the model so that the memory and communication effects can be entered directly as parameters in the model rather than derived indirectly from the model parameters. The resulting estimates can then be compared directly (and, if desirable, included in hypothesis tests). It is possible to produce, and it may be possible to estimate, a model which conditions on true values. In such a model the probability of a false positive would be distinguished from the probability of a false negative; this would make the model a better representation of reality. Alternatively it might be possible to implement a model which would incorporate all the elements of the interview-reinterview table rather than the summary provided by $g/(1-g)$.

10.1 The Impact of Interviewer Effects on the Study of the Relationships between Variables: Non-hierarchical Analysis - the Interviewer as a Term in Loglinear Analysis

In this section the analysis in chapter 3 section 3.2 is extended to cover the effect of the correlated response variance on the estimates of relationships, between variables. The data come from the Noise Annoyance Survey described in that section. Section 5.2.2 in chapter 5 describes a principal component analysis of the interviewer effects on the data.

The motivation for examining the relationships between annoyance, sensitivity and GHQ arose from earlier work on the pilot data (Tarnopolsky *et al.* 1978). In that analysis the single item 'being bothered' (item (1) from cluster 3 in table5.9 of section 5.2.2) was considered as a measure for annoyance. The three variables in the model were dichotomized and treated as nominal variables to facilitate a log-linear analysis. The analysis was presented in a path analytic framework following Goodman (1973), where annoyance was the main dependent variable, with the GHQ acting both as a response variable and a dependent variable. A measure of noise exposure was also considered but, of course, this was inappropriate in this instance as all of the interviews were concentrated in the high-noise zone.

In this section the early analysis is replicated, but this time the presence of an interviewer effect is taken into account. The results are shown in table 10.1.

Table 10.1 Relationships between annoyance, GHQ and sensitivity in the presence and absence of interviewer effect

(a) Logit analysis of proportion of highly annoyed in terms of sensitivity and GHQ

	With interviewer factor (INT)		Without interviewer factor	
	Estimate	S.E	Estimate	S.E
GM	1.8120	0.6202	0.4537	0.1597
SENS	1.2360	0.5528	1.0950	0.5205
GHQ	0.7294	0.3905	0.3759	0.3508
INT (2)	-2.2100	0.7306	-	-
INT (3)	-2.3760	0.7291	-	-
INT (4)	-0.5685	0.7720	-	-
INT (5)	-2.3050	0.7377	-	-
INT (6)	-1.4790	0.7298	-	-
INT (7)	0.0440	0.8731	-	-
INT (8)	-1.5680	0.7309	-	-
Residual sum of squares				
	20.8200	df = 18	53.3900	df = 25

(b) Logit analysis of proportion of high-GHQ scorers (more than 5 positive items) in terms of sensitivity

	With interviewer factor (INT)		Without interviewer factor	
	Estimate	S.E.	Estimate	S.E.
GM	-2.0910	0.5812	-1.368	0.1749
SENS	1.6650	0.4149	1.725	0.3899
INT (2)	0.0204	0.7966	-	-
INT (3)	1.6620	0.6698	-	-
INT (4)	0.1659	0.7930	-	-
INT (5)	1.1430	0.7096	-	-
INT (6)	1.0070	0.0737	-	-
INT (7)	0.6119	0.7307	-	-
INT (8)	1.6650	0.7426	-	-
Residual sum of squares				
	16.0500	df = 7	29.900	df = 14

Introducing an 'interviewer factor' reduces the total sum of squares by a significant amount ($P < 0.0001$). By examining the standardized beta coefficients in figure 10.1 we can see that the strength of the relationships between *GHQ*, and *Sensitivity* and *Annoyance* is actually increased when interviewer effects are taken into account. Incidentally, the model for *Annoyance* on *GHQ* and *Sensitivity* including an interviewer factor fits the data ($P > 0.1$), whereas the model without an interviewer factor present does not.

Figure 10.1 about here

This analysis suggests that in this case conclusions about the relationship between these factors are not substantially affected when interviewer effects are taken into account. The crude adjustment made in the analysis presented here tends to strengthen the evidence about the relationship. However, this result need not hold in all cases and further evidence about the effect of interviewer variance on estimates of the relationship is needed.

The results are encouraging. The analysis presented in this section demonstrates that without too much mathematical sophistication it is possible by means of modelling to adjust the analysis and in effect remove the impact of the interviewer effect from the analysis. It would appear that for the relationship examined here the distortion introduced by variability between interviewers has led to an attenuation of the coefficients obtained. This implies that the removal of the interviewer effects would strengthen the magnitude (and direction) of the relationships. This could, of course, lead to a failure to reject a null hypothesis when rejection would be appropriate, and where possible each relationship of interest should be tested separately for interviewer effect. In this paper we have considered only the simple additive model $y_{ij} = y'_{ij} + \alpha_i$. More complex models which include a component for the dispersion of each interviewer's results might provide some insight into the more general problems of attenuation.

10.2 Variance Component Analysis

In this section, the analysis of interviewer effect is extended by incorporating into substantive data analysis the differential impact of each of a group of interviewers on the responses obtained in two separate epidemiological surveys by means of variance component modelling. This strategy permits an evaluation of the impact of the interviewers on the interpretation of the linear models as well as allowing the use of characteristics of the interviewers to explain any variation introduced by the interviewers themselves.

10.2.1 Data Sources

The Physically Handicapped Survey (PHS)

The first study to be considered here is an experimental subsample designated for the exploration of interviewer effects in a longitudinal interview survey of services for the disabled. Full details of the sample design are given in Patrick (1981). The study was part of a large scale research programme into the health and care of the physically handicapped in the London Borough of Lambeth, under the sponsorship of the Department of Health and Social Security and with the assistance of the Special Trustees of St Thomas's hospital under the aegis of the Department of Community Medicine. The experimental project was partly funded by an ESRC award (HR 5971) to cover the first two years (waves) of the study. The subsample was concentrated in the southeastern part of Lambeth in an area covering four administrative wards containing 336 initial interviews randomly allocated across twelve interviewers. The analysis presented here is for the first wave of the survey. Two interviewers were excluded. One had entered the field experiment late in the first wave, and for another no interviewer level data were available as s/he was not originally expected to participate in the experiment. Thus ten interviewers are included in the analyses presented in section 4 with complete data across 137 respondents. Table 10.2 presents a summary of the response sets analysed.

Table 10.2: Sample allocation for the Physically Handicapped Survey (PHS)

Interviewer No.	1	2	3	4	5	6	7	8	9	10
No. of respondents	17	12	16	8	12	16	11	15	17	13

For the details of the original allocation scheme and analysis see Wiggins (1985). All of the interviewers were employed by Social and Community Planning Research (SCPR) and were regarded as 'panel' interviewers, i.e. interviewers who demonstrate a long term commitment to the organisation.

The choice of (individual level) variables for inclusion in the modelling was guided by analysis reported by Charlton given in Patrick (1981). The response variable, the Functional Limitations Profile, (FLP), provides a global measure of dysfunction (scale of 0-100) based on 135 items of daily living, originally developed as the Sickness Impact Profile by the Department of Health Services, University of Washington, USA. The original modelling strategy adopted by the PHS researchers was based on the analysis of (co)variance with three age groups. We decided to use respondent age defined as a continuous explanatory variable. Other explanatory variables included are sex (male/female), working status (at work/inactive), self assessment of health (ass. health), a 5-point rating defined as a continuous variable (5 indicates poor health), and a variable to reflect use of 'formal' services, in our case outpatient attendance at hospital (hsp. out) in the two weeks prior to interview (no/yes).

Interviewer level variables considered were the average number of calls made (ave. calls), age, sex, a supervisor rating of experience (5-point scale where 5 represented 'a lot') and an attitude score toward disability reflecting perceived differences between disabled and non-disabled people developed by Yuker et al (1970). A high score (scale 0-100) denoted a high level of tolerance towards the disabled.

The Aircraft Noise Survey (ANS)

The choice of (individual level) variables in the analysis was determined by the modelling included in O'Muircheartaigh and Wiggins (1981) and described in section 10.1. These models attempted to predict a tendency towards annoyance or bother with aircraft noise as a dichotomous item [for full details on the use of the original 'annoyance' scale see McKennell (1973)]. Two explanatory variables, psychiatric status (based on the 30-item General Health Questionnaire, GHQ, (Goldberg, (1979)) and sensitivity, an overall respondent evaluation of

his/her reactivity to aircraft noise, were defined as binary variables.

Five interviewer level variables were considered: average number of calls per workload (ave. calls), response rate for experimental assignment (resp. rt), interviewer gender (sex) and years of interviewing experience (yrs. exp.).

10.2.2 Analysis

The model is described in full in Wiggins, Longford, and O'Muircheartaigh (1992). An alternative formulation (of essentially the same models) is given in section 10.3. In both surveys analysed here the number of interviewers is small, and so complex modelling of the between interviewer variation is not meaningful. Therefore the number of interviewer-level variables in the fixed part and the number of variables in the random part has been limited.

PHS Analysis

Table 10.3 (column 1) gives the results of fitting the simplest one level model to the data from the first year of the PHS. This model ignores the presence of interviewers and is presented as a benchmark for the analysis. Only individual-level variables are included. There is a strong effect for work status and self-assessment of health; and weaker evidence of an effect for sex, and attendance at hospital outpatient services.

Table 10.3: Analysis of data from Physically Handicapped Survey; Functional Limitation Profile (FLP) score as dependent variable.

Explanatory Variables	(1) Fixed Effect (std. error)	(2) Fixed Effect (std. error)	(3) Fixed Effect (std. error)
Sex	2.36 (1.62)	2.33 (1.59)	2.28 (1.57)
Age	0.076 (0.065)	0.078 (0.064)	0.075 (0.69)
Work	4.05 (1.96)	4.15 (1.92)	4.08 (1.90)
Ass. Health	5.99 (0.93)	5.86 (0.90)	5.85 (0.89)
Hosp. Out	-3.19 (2.97)	-3.76 (2.91)	-3.74 (2.86)
Grand mean	-9.82 (0.79)	-9.10 (1.05)	-8.95 (1.03)
Random effects source	Sigma [Variance] (std. error)*	Sigma [Variance] (std. error)*	Sigma [Variance] (std. error)*
Respondent	9.21 [84.82]	8.92 [79.50]	8.81 [77.60]
Interviewer intercept	-	2.27 [5.56] (1.10)	2.22 [4.91] (1.01)
Slope for age	-	-	.091[.0083] (.078)
Deviance	-	994.58	992.42

* The standard error is the estimated standard error of sigma.

The model is extended in column 2 to allow a difference in intercept for each interviewer. Broadly speaking, the estimates of the fixed effects parameters are unchanged by the introduction of the random effect for interviewers. There is evidence of an interviewer effect.

A preliminary investigation of the effects of including interviewer-level variables in the analysis did not provide interesting results; this topic is pursued in our second application. For this example, we wanted to investigate the effect of adding a term to the random part of

the model. We chose age as an example since the prior literature suggests that response variance is related to age (O'Muircheartaigh, 1986, 1989) and that estimates of relationships involving age may thus be vulnerable to interviewer effect. We chose to examine this issue here even though the estimated variance component is not significant; our aim is to illustrate the available methodology. The model is the special case of the general model (1) in which only the coefficient for age has a positive variance in Σ.

The results are given in column 3. This leads to an improvement of the model, and the consequences are rather interesting. The fixed effects estimate of the regression coefficient on age is 0.078 but the square root of the corresponding variance (sigma) is much larger (0.091). Thus for a typical interviewer the slope on age is positive but the fitted variance of this slope is so large in comparison the slopes on age are predicted to be negative for a large proportion of the interviewers. Wiggins, Longford, and O'Muircheartaigh (1992) consider the residual effects due to age; they show that minimal variation (between interviewers) occurs for respondents around the age of 30 where there is no contribution of the interviewer variability to the total variance of an observation, but for both younger and older respondents there is a positive contribution to the total variance. The variance of an observation is a quadratic function of the age of the respondent and the minimum of this quadratic function occurs for age of about 40 years.

In terms of the impact of individual interviewers they show that the greater part of the variation arises from the contrast between interviewers 1, 2 and 5 on the one hand and 8 and 9 on the other. The results suggest that if interviewers 1, 2 and 5 were to carry out all the interviews, a strong positive association with age would appear, whereas if interviewers 8 and 9 only were used, there would be an apparent negative association with age.

Analyses of Aircraft Noise Survey data

The earlier analysis of these data, given in section 10.1, draws attention to the importance of introducing an interviewer factor into modelling relationships between substantive variables. A logit analysis was carried out in which a separate term was included for each interviewer;

this method is appropriate when dealing with a fixed set of interviewers.

Variance components analysis provides an approach to the analysis of these data where we wish to make inferences to a population of interviewers. Interviewers define the second level of the hierarchy. Here the response variable is binary (0, no annoyance; 1, annoyed) and the quasi-likelihood adaptation is used. A technical description may be found in Goldstein (1991). The variance for respondents has to be constrained to 1.0, by analogy with generalized linear models (GLIM) (see Nelder and Baker, 1981). Unlike the normal variance components, we do not have a value of the deviance (-2 loglikelihood).

Column 1 of table 10.4 presents the results of an analysis including only the fixed effects. Column 2 of table 10.4 shows the impact of including the interviewers as a random effect in the model. The results indicate that responses within an interviewer are rather highly correlated: $\rho = 0.408/1.408 = 0.29$. This intra-interviewer correlation is also equal to the variance components ratio - thus interviewer level variables could potentially explain up to 29% of the total respondent variation.

The five interviewer-level variables listed in section 10.2.1 were entered into the fixed part of the model one at a time. Their relative impact was judged in terms of the relative magnitude of their effect. Three had negligible impact - average number of calls per respondent, interviewer age, and interviewer sex. Response rate and years of experience did appear more convincing and are included in the final model given in column 3.

The effect of introducing these variables into the fixed part of the model is to separate out the effect of these particular aspects of interviewer performance from the overall variability introduced by the interviewers. The variables are introduced as individual-level variables; each individual in an interviewer's workload is allocated the score of the interviewer for a particular variable such as response rate. The effect of this is essentially to partition the overall interviewer effect into components due to (or explained by) particular aspects of the interviewer's performance and a residual component.

Table 10.4: Analyses of ANS data: annoyance (0, 1) as dependent variable

	(1)	(2)	(3)
Explanatory variables	Fixed effect (std. error)	Fixed effect (std. error)	Fixed effect (std. error)
Psychiatric status	0.372 (0.339)	0.562 (0.346)	0.586 (0.423)
Sensitivity	1.059 (0.468)	1.055 (0.485)	1.086 (0.496)
Grand mean	-0.147 (0.065)	0.108 (0.264)	0.071 (0.205)
Interviewer's response rate	-	-	-4.691 (2.918)
Interviewer's experience	-	-	-0.197 (0.089)
Random effects source	Sigma [Variance] (std. error)*	Sigma [Variance] (std. error)*	Sigma [Variance] (std. error)*
Respondent	1.000 [1.000]	1.000 [1.000]	1.000 [1.000]
Interviewer	-	0.639 [0.408] (0.218)	0.424 [0.180] (0.169)
Slope for psy	-	-	0.613 [0.376] (0.373)

* The standard error is the estimated standard error of sigma.

The results of this analysis can be seen by contrasting column 1 with the relevant parts of column 3. Two points are important. First, the substantive conclusions about the explanatory power of psychiatric status and sensitivity are virtually unchanged; the coefficients and their standard errors in columns 1 and 3 are practically identical. Second, the residual variance attributable to the interviewers is greatly decreased - the relative size of the variance component due to interviewers is reduced from 0.29 to 0.15 (the variance components ratio ρ is 0.180/1.180 = 0.15). In other words a substantial proportion of the interviewer variance can be explained in terms of achieved response rate and years of experience.

This variance component model corresponds to the logit analysis of covariance with

interviewers as the classifying factor. Introducing the interaction between this factor and an explanatory variable has a direct analogue in variance component analysis - it corresponds to having explanatory variables in the random part of the model. This allows the relationship between the response variable and the explanatory variables to vary from one interviewer to another. For example, column (3) of table 10.4 shows the effect of having 'psychiatric status' in the random part of the model.

The results indicate that those whose psychiatric status is positive are more likely to be annoyed by aircraft noise than those whose status is negative, even when their self-assessment of sensitivity to aircraft noise is taken into account. The relative odds calculated from column 2 suggest that for the high sensitivity group the probability of being annoyed is about 11% higher for the GHQ positive respondents (the expected proportions annoyed p are 0.848 and 0.767 respectively). The analysis in column 3 provides a check on the interpretation of this overall effect of psychiatric status shown in the fixed part of the model.

The value of sigma (0.613) provides information about the stability of the main effect of psychiatric status. The value suggests that the fixed effect may vary for different interviewers, but it should be borne in mind that the sigma value is an estimate with estimated standard error of 0.373.

10.2.4 Conclusion

Historically, the investigation of interviewer effect concentrated on univariate analysis (e.g. Kish, 1962). More recently, advances in statistical computing (e.g. VARCL, Longford, 1988 and ML3, Goldstein, 1987) permit investigators to model relationships between variables while at the same time allowing for the differential impact of interviewer performance and different interviewer characteristics. The two illustrations presented in this section are based on studies in which the allocation of individual respondents to interviewers is randomised and in which the primary objective was the construction and evaluation of linear models among variables. Although the number of different interviewer characteristics was limited, the results show a convincing potential for subsequent methodological investigation.

The first application (PHS) demonstrates the method for a quantitative response variable. While the basic analysis provides evidence for the presence of a substantial interviewer effect, the simple interviewer effect model does not seriously contaminate the interpretation of the fixed part of the model. Extending the analysis to allow for variability in the fixed part of the model (in particular allowing for a random coefficient for respondent age) we see that if interviewing were confined to particular subsets of interviewers the estimate of the regression coefficient in the fixed part of the model could be seriously biased.

The second application (ANS) presents an illustration for the quasi-likelihood adaptation of the estimation procedure for a binary response variable. Again, the simple variance components model confirms the presence of a substantial interviewer effect (29% of variance can be attributed to the interviewers themselves) whilst not disturbing any interpretation of the fixed part of the model. Attempts to explain the presence of interviewer variance by including interviewer characteristics in the model suggest that variation in interviewer response rate and experience can account for about 50% of the total variability introduced into the responses by the interviewers.

On balance these illustrations of variance component models provide a potentially valuable approach to the analysis of the impact of variability among interviewers on the results of sample surveys. By directly incorporating the presence of interviewers and their characteristics into the modelling process the concerns of the methodologist and of the data analyst are integrated into a single unified approach.

10.3 The British Household Panel Survey[1]

Whereas it has been possible for many years to carry out a simultaneous analysis of interviewer and cluster effects for sample means and other simple statistics[2], it is only recently

[1] Another version of the material in this section may be found in O'Muircheartaigh and Campanelli (1998)

[2] Technically, means and proportions estimated from survey data are ratio estimates as there is uncontrolled variation in the sample size. For the BHPS the selection of PSUs with probability proportional to size and equal probabilities overall, this variation is fairly tightly controlled.

that software has become available to estimate interviewer and cluster effects simultaneously while incorporating these effects directly into a substantive model of interest. This is possible through the use of a cross-classified multilevel model using the software package MLn (Rasbash et al, 1995); alternative programs for multilevel analysis are VARCL (Longford, 1988) and HLM (Bryk et al, 1986). Section 10.2 gives two early applications.

This section compares the relative impact of interviewer effects and sample design effects on survey precision by making use of an interpenetrated PSU/interviewer experiment which was designed by the authors for implementation in the second wave of the British Household Panel Study (BHPS). The data themselves and the results of the application of the standard methods are given in section 3.3 of chapter 3. The models used here are described in section 10.3.1; section 10.3.2 explores the results over all BHPS variables and illustrates the use of a multilevel (hierarchical) approach in which the interviewer and sample design effects are estimated simultaneously while being incorporated in a substantive model of interest. Finally, section 10.3.3 summarises and discusses the findings and their implications for survey research practice.

10.3.1 Cross-Classified Multi-level Models

An alternative conceptualization of the analysis is as a multi-level (hierarchical) model in which the interviewer, PSU, and geographic pool are hierarchical partitions and the terms corresponding to them in the model are considered to be random effects. It is only recently that cross-classified multilevel analysis has become feasible (see Goldstein, 1995, Rasbash et al, 1995); the design is implemented in MLn by viewing one member of the cross-classification as an additional level above the other. A basic multilevel variance components model to capture the interviewer by PSU cross-classification within geographic pool can be defined as follows:

$$y_{i(jk)l} = \alpha + \beta x_{i(jk)l} + u_j + u_k + u_l + e_{i(jk)l} \qquad (10.8)$$

for the i th survey element, within the j th PSU crossed by the k th interviewer, within the l

th geographic pool, where $y_{i(jk)l}$ is a function of an appropriate constant (α), an explanatory variable x and its associated coefficient β, and an individual error term ($e_{i(jk)l}$). Here u_j is a random departure due to PSU j, u_k is a random departure due to interviewer k, and u_l is the random departure due to geographic pool l. Each of these terms and $e_{i(jk)l}$ are random quantities whose means are assumed to be equal to zero. In cases where the dependent variable is a dichotomy, $y_{i(jk)l}$ would be replaced in (10.8) by log ($\pi_{i(jk)l}$/1-$\pi_{i(jk)l}$), where

$$\pi_{i(jk)l} = \frac{\exp(\alpha + \beta x_{i(jk)l} + u_j + u_k + u_l)}{1 + \exp(\alpha + \beta x_{i(jk)l} + u_j + u_k + u_l)}$$

When the dependent variable is continuous, ρ can be calculated directly from the variance estimates in a variance components model (e.g., interviewer variance divided by total variance). When the dependent variable is dichotomous, the variance components are given on the logistic scale and a more complex computation is required. We generate random normal deviates with variance given by the component estimate. These deviates are then transformed (taking the anti-logit) and the variance of these transformed values is calculated directly to give the numerator for ρ.

The treatment of the interviewer and PSU effects as random effects rather than as fixed effects (more common in the survey sampling literature) postulates a `superpopulation' of interviewers from which the interviewers used in the study were drawn and an infinitely large population of PSUs. In the case of interviewers we can consider the inference as being made to the population of potential interviewers from whom the survey interviewers were drawn. For the PSUs the assumption involves essentially ignoring a small finite population correction (see, for example, Kalton, 1979). As we are interested in the relative magnitudes of the components of variance due to the interviewers and the sample design under the essential survey conditions this treatment will not affect our conclusions materially.

An added advantage of multilevel modelling in general (cf Hox, de Leeuw, and Kreft, 1991; Wiggins, Longford and O'Muircheartaigh, 1992) is the facility to incorporate covariates directly into the analysis. For our work we will be able to examine such factors as interviewer

age, gender, length of service, status, and whether the same interviewer was present for both Wave 1 and Wave 2 of the panel survey. We can also include respondent characteristics. Area level characteristics based on a match to census small area statistics could be added in due course. Single level linear models have of course been used to analyze survey data. Such non-hierarchical models ignore the way in which the clustering in the sample design and the clustering of responses generated by the interviewers may affect the variance-covariance structure of the observations.

10.3.2 Results

For illustration, we include three MLn models, one for each of the main types of variables: interviewer check items, facts, and attitudes. These are shown in tables 10.5-10.7 respectively. We have also shown the corresponding non-hierarchical (single level) model to discover whether our substantive conclusions will be affected when we incorporate the data structure appropriately in the analysis.

The variable modelled in table 10.5 is a binary subcategory indicating whether children were present during the demographics section of the interview, as noted by the interviewer. From the hierarchical analyses of variance, the estimated ρ values for this *children present* subcategory were $\rho_i = 0.171$ and $\rho_s = 0.062$ (n = 725).

Table 10.5: Multilevel Logistic Regression Model of Interviewer Check Item: Children Present

	Model 1		Model 2		Model 3	
EXPLANATORY VARIABLES	FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)		FIXEDEFFECT (Std. error)	
	Non-hier	Hier	Non-hier	Hier	Non-hier	Hier
Grand mean	-1.05 (0.08)	-1.05 (0.14)	-3.24 (0.37)	-3.30 (0.41)	-5.42 (0.94)	-5.49 (1.27)
No. of children in HH	-	-	1.20 (0.10)	1.23 (0.11)	1.23 (0.10)	1.25 (0.11)
Respondent gender (F)	-	-	0.62 (0.21)	0.59 (0.22)	0.62 (0.21)	0.60 (0.22)
Interviewer gender (F)	-	-	-	-	1.11 (0.43)	1.14 (0.62)
RANDOM EFFECTS SOURCE	VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)	
Respondent	-	1	-	1	-	1
PSU	-	0.09 (0.12)	-	0.08 (0.17)	-	0.08 (0.17)
Interviewer	-	0.49 (0.20)#	-	0.89 (0.32)#	-	0.81 (0.31)#

Significant random parameters based on contrast test.

The hierarchical version of Model 1 is a basic variance components model showing the cross-classification of PSU and interviewer. Although the estimated standard errors of the random parameters are included in the table, the significance of the random parameters is based on a contrast test.[3] We found significant variation between interviewers but not between PSUs. In the model the estimate for variation between geographic pools for this variable was zero; this was not of course the case for all variables. Parameters close to zero are often constrained to zero by the MLn programme; in this case the parameter remains zero even

[3] As the distribution of the standard errors for the random parameters may depart considerably from normality, especially in small samples, a better option is to use a specific contrast on the random parameters.

when employing the `second order estimation procedure'.[4] In the standard formulation of the model the individual variation is assumed to have a binomial distribution and is constrained to 1.[5]

In model 2, we have included the individual level explanatory variable, number of children in household, as it seems desirable to control for any systematic differences among interviewers in the composition of their workloads; an interviewer whose interviews take place in households without children would be expected to differ on this item from those interviewers whose workloads contained a large number of households with children. This control variable has a significant coefficient in the hierarchical model. For fixed effects significance may be judged by comparing the estimate with its standard error in the usual way.

Also included is the individual level explanatory variable, respondent's gender. We expected that the presence of children during the interview would be a function of the respondent's gender, with women respondents being more likely to have children with them than male respondents. As can be seen by the values in table 10.5, this expectation was confirmed.

It is interesting to note that the random coefficient for interviewers in the hierarchical version of the Model 2 increases in comparison to Model 1. This suggests that it is not haphazard variation in interviewer workloads that explains this interviewer variability, but rather that the variation among interviewers in recording the presence of children is greater when opportunity (ie children in household) is taken into account as well as respondent's gender. The basic conclusion which can be drawn from Model 2, however, is the same for both the hierarchical and non-hierarchical versions of the model.

We then proceeded to add in several interviewer explanatory variables. These included interviewer age, gender, status (whether basic interviewer, supervisor, or area manager), and

---

[4] In the estimation of random parameters in a logistic regression, MLn uses a weighted generalised least squares estimation procedure which requires the quantities to be estimated to be in the linear part of the model. A series expansion is used to approximate a linear form. Simulation and theory have suggested that the first order estimation procedures can lead to an underestimation of the parameters. In many models the underestimation is negligible. However, in some models where predicted probabilities are extreme, or where there are few level 1 units per level 2 unit, underestimation can be severe. An option exists in MLn which allows one to select a second order estimation procedure. This procedure, however, is less computationally robust. See Woodhouse (1995) for a full description of this matter.

[5] The validity of this assumption can be tested in MLn by relaxing this constraint.

years with the company. Also included was a measure of whether the same interviewer had visited the household for last year's interview. Of these various characteristics, only interviewer gender is considered in Model 3. It was clearly significant in the non-hierarchical model and only approached significance in the hierarchical one. It is interesting to note that in this case, different conclusions might have been reached depending on which model was considered. We also investigated the possibility of an interaction between interviewer gender and respondent gender. This coefficient was not significant under either version of the model.

There are at least two possible explanations for the correlated interviewer effect in this case. First, there is quite likely a difference in the ability of interviewers to arrange the circumstances of the interview so that the respondent is alone at the time - flexibility in making appointments, degree to which the interviewer emphasises the need for an undisturbed setting for the interview, etc. There is also the possibility that most of the between-interviewer variability is due to differences in the extent to which, or the circumstances in which, interviewers record the presence of children; one source of variation could be in the definition of others being `present'.

The key contrast here is between the message we would obtain from ρ_i and ρ_s and the message from the multilevel analysis. With the former we would be concerned that the standard analysis would give spurious significance to the relationships estimated. In this case at least, however, interviewer effect - though present for the dependent variable - does not affect the substantive analysis.

Table 10.6: Multilevel Logistic Regression Model of Newspaper Readership: Reads the *Independent*

	Model 4		Model 5		Model 6	
EXPLANATORY VARIABLES	FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)	
	Non-hier	Hier	Non-hier	Hier	Non-hier	Hier
Grand mean	-3.04 (0.13)	-2.99 (0.30)	-1.70 (0.35)	-1.94 (0.45)	-2.99 (0.67)	-3.19 (0.90)
Respondent's age	-	-	-0.03 (0.01)	-0.03 (0.01)	-0.04 (0.01)	-0.03 (0.01)
Whether same interviewer as last year	-	-	-	-	0.21 (0.28)	0.63 (0.34)
Interviewer status Whether regular interviewer (compared to area manager)	-	-	-	-	1.35 (0.60)	1.06 (0.84)
Interviewer status Whether supervisor interviewer (compared to area manager)	-	-	-	-	2.25 (0.76)	2.23 (1.25)
RANDOM EFFECTS SOURCE	VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)	
Respondent		1		1		1
PSU		1.55 (0.64)#		1.48 (0.63)#		1.59 (0.66)#
Interviewer		1.97 (0.71)#		1.78 (0.68)#		1.67 (0.67)#

Significant random parameters based on contrast test.

Table 10.6 deals with one of the respondent level factual items, newspaper readership. The variable modelled is a binary subcategory indicating whether or not the respondent typically

reads the *Independent.* From the hierarchical analyses of variance, the estimated ρ values for this readership subcategory were $\rho_i = 0.129$ and $\rho_s = 0.106$ (n = 1,268).

Unlike the variance components model shown for the interviewer check item (see Model 1), the basic variance components model given in Model 4 shows significant variation between PSUs as well as between interviewers. For this also there was no significant variation between geographic pools.

In model 5, we have included the individual level explanatory variable, respondent's age. Several other explanatory variables had also been explored in both the hierarchical and non-hierarchical versions of the model (e.g. gender, social class, political party identification, and income) but only respondent's age was significant. With this addition, the interviewer random variation is reduced slightly and the PSU random variation remains essentially the same.

Of the various interviewer explanatory variables we considered, two approached significance in the hierarchical version of Model 6. These were the binary variable for whether the same interviewer had visited the household for last year's interview (interviewer continuity) and one of the two dummy variables modelling the 3 category interviewer status variable (regular interviewer, supervisor, area manager). Here we can see that the interviewer variance component is again slightly reduced.

Interestingly we would have had a very different interpretation of which interviewer characteristics are having a significant impact if we had only run the non-hierarchical model. With the non-hierarchical model, the interviewer continuity variable was clearly not significant and the two interviewer status variables were clearly significant. In addition, (although not shown in table 10.6), interviewer's age approached significance. Middle-aged interviewers were more likely than elderly ones to record respondents as readers of the *Independent.*

Table 10.7: Multilevel Logistic Regression Model: Likely Have More Children?

	Model 7		Model 8		Model 9	
EXPLANATORY VARIABLES	FIXED EFFECT (Std. error)		FIXED EFFECT (Std. error)		FIXEDEFFECT (Std. error)	
	Non-hier	Hier	Non-hier	Hier	Non-hier	Hier
Grand mean	-0.39 (0.06)	-0.44 (0.11)	7.73 (0.46)	7.59 (0.46)	8.81 (0.60)	7.39 (0.48)
No. children in HH	-	-	-0.85 (0.09)	-0.83 (0.10)	-0.86 (0.10)	-0.84 (0.10)
R's gender (F)	-	-	-0.65 (0.19)	-0.63 (0.19)	-0.64 (0.19)	-0.62 (0.19)
R's age	-	-	-0.24 (0.01)	-0.23 (0.01)	-0.24 (0.01)	-0.24 (0.01)
Interviewer years w/ co.	-	-	-	-	0.042 (0.020)	0.043 (0.027)
RANDOM EFFECTS SOURCE	VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)		VARIANCE COMPONENT (Std. error)	
Respondent	-	1	-	1	-	1
PSU	-	0.15 (0.09)	-	0.00 (0.00)	-	0.00 (0.00)
Interviewer	-	0.22 (0.10)#	-	0.38 (0.16)#	-	0.34 (0.15)#

Significant random parameters based on contrast test.

Table 10.7 presents a behavioral intention item looking at whether or not the respondent expects to have any more children. As this is a subjective assessment, the question has been classified in the attitude category for our analysis. From the hierarchical analyses of variance, the estimated ρ values for this item were $\rho_i = 0.075$ and $\rho_s = 0.048$ (n = 1,177).

As was the case for the variance components model under Model 1, Model 7 shows significant variation between interviewers, possible variation between PSUs, but not among

geographic pools.

In Model 8, we have included the individual level explanatory variables, number of children in the household, respondent's gender, and respondent's age. Each of these is highly significant in both the hierarchical and non-hierarchical versions of the model. With the addition of these explanatory variables in the hierarchical model, random variation due to PSUs goes to zero and random variation due to interviewers increases. The disappearance of the PSU effect may mean that the characteristics that led to the possible PSU effect have been adequately specified in the substantive model. Again, this suggests that it is not haphazard variation in interviewer workloads that is contributing to interviewer variability, but rather that there is variation among interviewers in their measurement of people's intentions to have more children.

In the non-hierarchical version of Model 9, interviewer experience is a significant predictor with more experienced interviewers being more likely to record a `yes' to the *more children* question than inexperienced interviewers. Although not shown, in the non-hierarchical model the interviewer continuity variable approached statistical significance. When the same interviewer returned on the second wave of the survey he/she was less likely to record a `yes' to the `more children' than a different interviewer. This finding, however, does not hold for the hierarchical model.

Perhaps the most important point to note here is that, despite the strong interviewer effect, the substantive description represented by the substantive fixed part of the model is unaffected by the interviewers (at least not affected differentially). However, there are differences in the conclusions about the impact of interviewer characteristics depending on whether an interviewer variance term is explicitly included.

In addition to these examples above, we conducted a further exploration of the impact of the extra-role characteristics of the interviewers (Sudman and Bradburn, 1974) on model conclusions. For each of the different types of item (attitudes, facts, quasi-facts, and interviewer checks), a sample of variables was drawn from among those shown to have highly

significant interviewer variability. Across the four categories, 26 items were drawn from 84. A cross-classified multi-level analysis (interviewer by PSU) was conducted on each one of these with the interviewer characteristics as the explanatory variables. These included interviewer age, gender, status, years with the company and an indicator of interviewer continuity over time. Of the 26 models considered, interviewer age was significant in 7 of the 26 cases (27%). The comparable percentages of significant effects that were found for the other interviewer characteristics were as follows: interviewer continuity (12%), gender (8%), interviewer status (8%), and years with the company (4%). Although such data should be treated with caution, it may indicate that interviewer age is a general predictor of some of the interviewer variability on the high variability items. Freeman and Butler (1976), for example, found age and gender to be significant predictors of interviewer variance. Collins and Butcher (1982) also investigated the explanatory power of several characteristics of interviewers. Their strongest evidence was for an age effect.

Again we saw differences depending on whether a hierarchical or non-hierarchical model was used. The comparable figures for the non-hierarchical models were: age significant in 27% of cases, interviewer continuity in 15%, gender in 12%, interviewer status in 35%, and years with the company in 15%. In 11 of the 26 models, different conclusions about the effects of interviewer characteristics on substantive results would have been reached, depending on whether an interviewer variance term was explicitly included in the model.

10.3.3 Summary and discussion

The assumption underlying most statistical software - that the observations are independent and identically distributed (*iid*) - is certainly not appropriate for most sample survey data. Variances computed on this assumption do not take into account the effects of survey design (eg inflation due to clustering) and execution (eg inflation due to correlated interviewer effects).

Using software developed for multilevel analysis (hierarchical modelling) we presented an

alternative framework within which to consider the sample design and interviewer effects by incorporating them directly into substantive models of interest. For illustration we chose three binary items - an interviewer check item on *whether children were present during the interview*, a behavioural item, *readership of the Independent*, and a subjective item, *whether respondents thought it was likely that they would have another child.* For each of these items, we found a significant interviewer effect, which persisted when we controlled for inequalities in the interviewers' workloads and various extra-role characteristics of the interviewers. For other items not presented here we found situations where interviewer characteristics did help to explain the interviewer effects. In addition, we found that conclusions about the influence of the various extra-role characteristics would have differed in many cases if we had used only the standard non-hierarchical model rather than a hierarchical one.

In later work we hope to explore further the factors that might provide an explanation of the variance components. From a modelling standpoint the issue is one of specifying appropriately the underlying factors in the substantive models of interest. From a sample survey standpoint the issue is that of incorporating in the analysis a recognition of the special features of the sample design and survey execution that make a particular data set deviate from *iid.* Multilevel models have a natural congruence with many important aspects of the survey situation; both the sample design and the fieldwork implementation can be described appropriately as introducing *hierarchical levels* into the data and thus multilevel analysis provides a framework that makes it possible to include both substantive and design factors in the same analysis. Furthermore they provide a framework within which the possibilities of extracting substantive information from response error can be explored.

Chapter 11 - AN OVERVIEW

This chapter synthesizes the ideas expressed in the earlier chapters. It is not intended as a summary of the earlier chapters but as a review of the conceptualization.

11.1 The Survey Sampling Perspective

The different literatures that spawned representations of measurement errors in surveys are as diverse as the backgrounds and motivations of the various subgroups that used the survey as a research vehicle. They range from social reformers, whose interest was in collecting data with which to impress on a reluctant and sometimes antagonistic world the case for reform, through advertisers and manufacturers who wished to sell their products, through academic researchers whose interest was in the measurement process itself, to market researchers for whom the needs of the client (and the need to get paid) were the primary motivators. Partly due to their different agendas and partly due to their different disciplinary backgrounds the ways in which they conceptualized the notion of error were different. Chapter 1 discusses these issues at more length.

My first exposure to the ideas of measurement error came in the context of sampling error. In this framework the survey estimate is seen as a sample-based statistic that estimates a population-based parameter. The inferential framework is Neyman's repeated sampling inference where the variability in the estimate (and consequent potential discrepancies between the estimate and the parameter being estimated) is described as sampling error. This is usually measured by the variance of the sampling distribution of possible outcomes of the sampling procedure (thus Neyman's inference is sometimes termed a *procedural inference*) . This framework easily generalizes to any number of factors. The most notable expansion was that by Mahalanobis, who considered the whole survey activity in terms of replications of the basic design, and therefore incorporated measurement errors generated by various elements of the data collection into a *total variance*.

Statisticians when faced with a variance that has many sources automatically turn to the

analysis of variance [ANOVA] as a means of partitioning and understanding it. Survey statisticians were no different; Yates' paper on sample designs and sampling variances in 1946 was couched entirely in terms of ANOVA models (though this approach to sampling did not gain favour again for some 30 years or more). ANOVA was also a basic tool for survey statisticians and it was the method of choice for analyzing the impact of different factors on survey data, and consequently dictated the form of the conceptualization of the total variance (or perhaps vice versa). Kish's 1962 paper gives the formal application of this model to the impact of interviewers on individual survey responses, though Durbin, Stuart, Moser, Kendall, and Gales had all used some type of ANOVA models in assessing classes of interviewers.. A somewhat different representation of the phenomenon, but with effectively the same estimators, was presented by Fellegi in 1964 in an extension of the model for measurement error being used by the US Bureau of the Census.

Under the ANOVA model the total variance of the estimate can be partitioned into four components, as presented in chapter 4. The two sources of error are *sampling* and *measurement* and within each there are two types, *uncorrelated* errors and *correlated* errors. As a shorthand we use SSV for the simple (or uncorrelated) sampling variance, SRV for the simple (or uncorrelated) response variance, CSV for the correlated sampling variance, and CRV for the correlated response variance. The SSV is the variance that the estimate would have if a simple random sample of observations were selected from the population and there were no measurement error. The CSV is the additional variance brought about by having to select clustered samples, so that additional elements (individuals) from within the same cluster introduce artificially similar elements into the sample compared to what we would get if we were to select more elements completely at random from the population. Another (almost accurate) way of expressing this would be to say that additional elements from clusters already represented in the sample are less informative than elements from clusters not represented. Thus the SSV will misrepresent the variability in the estimate due to sampling.

On the measurement side, the SRV is a measure of the basic unreliability of the data. It

incorporates the impact of all the haphazard and situational factors that influence the survey response (though not the systematic biases that affect all the elements or all the situations). The CRV is the additional variance brought about by factors that influence in a uniform way sets of elements. The most obvious example - and the one that has generated the most literature - is the effect of the interviewer. If an interviewer has a systematic effect on all the respondents s/he interviews, then this will make the data seem more homogeneous than the underlying correctly measured values warrant, and the variance estimated without taking that into account will underestimate the actual variance of the estimate. This is a similar impact to that of clustering in sample design and it can be represented in an ANOVA model in a similar way.

[It may be worth reiterating here briefly the idea of *essential survey conditions* introduced by Hansen, Hurwitz, and Bershad (1961) - the Census Bureau Model. Following a conceptualization that is in harmony with Neyman's, the Census Bureau Model postulates a hypothetically repeatable measurement process where the same individual can be measured a(n infinite) number of times. This repetition itself generates a distribution of measurement errors and it is the variance of this distribution that is measured by the SRV. The essential survey conditions are the factors or values considered fixed during the whole set of hypothetical repetitions. Any systematic effect of these conditions is not measured by our variances and constitutes bias under this model.]

The SSV and the SRV are the basic building blocks in the total variance. The other components represent the effects of our interventions - whether by design decisions for the sample or by implementation decisions for the fieldwork (data collection). The SSV represents the true (and, in a sense, irreducible) variability in the population. We can change the impact that this variability has on our estimates by judicious choice of sample design; design decisions that increase the variance will usually be warranted by cost considerations. The net impact of our decisions will be embodied in the CSV. [It is in principle possible that the net effect of sample design could be to make the TSV (total sampling variance) less than the SSV, but in practice for demographic surveys this is never the case.] In this dissertation the sampling variance is discussed primarily to give a frame of reference for the assessment

of the response variance.

The SRV represents the basic variability of the measurement process given the essential survey conditions. The model on which the SRV is based assumes essentially that there is no structural factor in the survey design that would affect the additivity of the many subcomponents of the SRV and that it can therefore be estimated using a simple within-sample variance estimate.

In practice, of course, this is not the case. Many of the components of a survey execution plan involve a non-random structure and non-random groupings. Interviewers interview many respondents - the economies of scale dictate that once an interviewer is trained it is desirable to spread the cost of training that interviewer over as many respondents as possible; supervisors are responsible for many interviewers; coders of open-ended responses will code a large number of questionnaires; question formats are repeated within batteries of questions in a questionnaire. To the extent that each (or any) of these factors imposes a common effect on the elements to which it is applied, the measurement variance is changed from the SRV to something that incorporates that common effect, either as an additional component in the ANOVA model of total variance or as a covariance in the correlation/covariance model. The aspect of these structures and groupings most relevant to this discussion is that they are the result of decisions made in planning the survey instrument and the fieldwork; critically, they are under our control as survey designers.

Constructing a plausible case for the existence of correlated response errors does not establish that these errors have a significant role in determining the precision of survey estimates. The early research in this area was directed at this question, and the early studies could be regarded as exploratory; their first purpose was to establish the existence of correlated errors of magnitude sufficient to affect survey conclusions. The early studies of interviewer effect were essentially *identification studies*. The two pioneering studies (Kish, 1962, and Fellegi, 1964) took somewhat different approaches. Kish explicitly addressed the issue of establishing the existence of "interviewer effect" using a frequentist argument involving the expected number of such effects significantly different from zero found in the analysis of some batteries

of attitudinal items. Fellegi simply presented the results for a set of items from the Canadian Census of Population without an explicit consideration of the issue. An earlier study worth noting is Gales and Kendall (1957) though it considered differences among classes of interviewers rather than among individual interviewers (what are now called "house effects" in market and social research).

Among survey practitioners there remained some question about how much importance should be attached to an effect that did not manifest itself universally, but whose magnitude when it did manifest itself was such that it could dominate all other measurable sources of error. The general approach was to search for a categorization of variables that could predict when an interviewer effect would arise. The early attempts at categorizing variables were theory-free; having estimated interviewer effects for a large number of variables an explanation was sought for the cases where the estimated interviewer effects were largest. Such studies of course are vulnerable to the over-interpretation of random outliers and the corresponding neglect of real effects not manifested in a particular study - the classic problem of false positives and false negatives. One exception is Kish's 1962 study, in which questionnaire items were classified a priori by predicted sensitivity to interviewer effects. Though the empirical results did not support the classification, the study is nevertheless important in showing an awareness of the need for a theory-driven approach.

This uncertainty about the nature of interviewer effect has remained a problem in its interpretation. The empirical results obtained from studies which calculate values of ρ for sets of variables have two rival interpretations: either all variables could be affected by interviewer effect, and the variation in estimated ρ could be caused by estimation error, or, only some variables might be affected, and the variation is a reflection of that different sensitivity. The two interpretations have different implications in terms both of survey design and survey analysis. Without replication, or else access to *true values* for the element-level data, the two interpretations are necessarily confounded. Two of the sets of studies described in this dissertation address this specific issue.

In the World Fertility Survey (WFS) Response Errors Project the same (or at least a

standardized) questionnaire was used in each country. Thus the studies provide replication of the same set of variables in different contexts; indeed the socio-cultural differences among the countries provide an even sterner test of replicability than would normally be demanded. There is an imposed similarity in the essential survey conditions in that central WFS staff and procedures were used in setting up the fieldwork and training the interviewers in all the countries. Chapter 4 presents the results of these analyses.

The results (values of ρ) show a striking similarity of pattern for Lesotho and Peru. Remember that Lesotho and Peru have different languages, different cultures, a different religious composition, and are on different continents. Nevertheless when we look at the variables with statistically significant values of the intra-interviewer coefficient we find that the variables not only coincide exactly, but that the order of sensitivity is very similar, despite the fact that all the estimates of intra-interviewer effect are themselves of course subject to estimation error.

At least for this set of variables it appears to be convincing that certain identifiable variables are more sensitive to systematic interviewer effect than others. This makes it possible to begin to examine the dimension(s) along which this sensitivity may vary. It also suggests that there is a judgment to be made about the loss function involved in optimizing a survey design in terms of the total variance. As with sample design, the optimum design for one variable may be quite different from that for another variable. Interestingly - and perhaps generalizably - the variables of most interest may well be the most sensitive.

[These results should at least raise in our minds the question as to how we present survey estimates, and how we interpret them. When we present confidence intervals in the Neyman framework we find it difficult to reach any strong conclusions if we treat all the values in the confidence interval as being in some sense "equally plausible". Though it would be foolish to treat every point estimate as a parameter value, it may be almost equally unwise not to accept that the point estimate is the most appropriate value to use for some interpretative purposes. The results are a useful counterpoint to the necessarily cautious interpretation of variability in estimates that statisticians normally put forward. By training, statisticians -

especially survey statisticians - tend to present a confidence interval as the most appropriate information about a parameter. A subject matter analyst will tend to concentrate on the point estimate, and discuss its importance implicitly assuming that it is correct.

The results obtained for sampling variance for WFS reinforce this thinking (see Verma, Scott, and O'Muircheartaigh, 1980). Again the consistency across countries of the estimates demonstrates that treating the estimates as parameter values is on balance less misleading than treating the whole interval as equally plausible; there are however obvious dangers to carrying this argument to its logical conclusion.]

The second set of analyses that addresses this issue is the British Household Panel Survey (BHPS) data set (especially in chapter 3). The strength of the data set is in the number of variables that can be analyzed, allowing patterns of variation to emerge by averaging across variables of the same (or similar) types rather than across countries as for the WFS data. The data set contains 820 variables. This means that for the first time (at least among published reports) it is possible to consider interviewer effects in the same way that sampling errors have been considered for decades.

There has been a whole industry dedicated to the calculation of the impact of sample design on the precision of survey estimates; the notion of the *design effect* (*deff*) has generated enormous amounts of literature and attention. Its widespread acceptance has come not from evidence that *all* variables are affected by intracluster homogeneity, but first from its plausibility as a concept, and second from strong evidence that *on average*, or *in general*, variables in surveys show clear evidence of this phenomenon. Where for a particular variable no effect is found, this is first considered to be an estimation error (sampling error in the estimate of the effect) and only in special cases thought to be that the variable is not in fact spatially clustered. Indeed it is the recommended policy for modeling sampling errors that average effects be taken rather than relying on the point estimate in a particular case. For estimates of sampling error that are themselves subject, of course, to sampling error, this is clearly a sensible approach. It is both more robust and under certain assumptions more precise than a direct estimate for each variable.

It is rare now that anyone would suggest that design effects are not important. The conventional view is bolstered by studies in particular contexts such as achievement tests in school surveys, where very large effects are found and thus it is extremely rare that a negative or even negligible value of intracluster correlation is found. The evidence for design effects is of course extensive as all that is required to obtain an estimate is a measurable sample design (ie a design with at least two sampling units in each stratum).

The situation with regard to intrainterviewer correlation has always been very different. The number of studies is small and the number of variables considered in each is small. This makes an analysis, or indeed a body of literature, similar to that for sampling errors impossible to achieve. For the BHPS data, however, similar analyses can be carried out for sampling variance and for response variance.

This is the traditional sample survey approach and includes consideration of the *design effect* and the *interviewer effect* following the ANOVA and Census Bureau models. The analysis considers only the estimation of means and proportions and their standard errors. Across the 820 variables in the study, there was evidence of a significant impact of both the population clustering and the clustering of individuals in interviewer workloads. The intraclass correlation coefficient, ρ, was used as the measure of homogeneity. We found that sample design effects and interviewer effects were comparable in impact, with overall inflation of the variance as great as five times the unadjusted estimate. The median effect across the 820 variables was an 80% increase in the variance.

If it were true in general (and it is in this survey) that interviewers have an impact on the precision of estimators comparable to that of the clustering in the sample design, then equal attention should be paid to the design of surveys to mitigate the effect. The large values of ρ_i on particular items and the fact that ρ_i is of the same order of magnitude as ρ_s suggests that survey organisations should attempt to incorporate measurement of ρ_i into their designs. If the necessary modifications of the survey design are too expensive to allow this, organisations should at least try to minimise its impact; this could be accomplished by reducing interviewers' workloads. Current practice tends to favour smaller dedicated interviewer forces with large

assignments; in the presence of substantial interviewer effects this is a misguided policy.

The large number of variables also makes it possible to consider categories of variables separately while still having a sufficient number of variables within category to draw reasonably robust conclusions about sensitivity to interviewer effect. The categories were: factual items, attitude items, self-completion items, and interviewer check items. Though these might be expected to behave differently from each other it is worth noting that they are all socio-economic or socio-demographic variables and thus do not span the full spectrum of survey questions.

The results were fascinating. The pattern of estimates was strikingly similar for each category, and the order of magnitude of the estimated coefficients was comparable across categories, though interviewer check items (items that the interviewer completed from observation) were somewhat more sensitive than the other categories.

One further result, though difficult to interpret, is potentially very exciting. There was a correlation, across all the variables (that is, treating the variables as the entities being correlated) between the intrainterviewer correlation coefficient (the *interviewer effect*) and the intracluster correlation coefficient (the *clustering effect*). I interpret this to mean that the extent of spatial/demographic/cultural homogeneity is positively correlated with the level of sensitivity to interviewer distortion. This result has already generated considerable scepticism; it is appropriate that it should do so.

The first problem is in understanding what is being described. The unit of observation in calculating this correlation is the variable (a question or a category in a question). A positive correlation therefore means that variables that have a relatively high value of the intracluster correlation coefficient will also, on average, have a relatively high value of the intrainterviewer correlation coefficient. Thus as with any other product-moment correlation coefficient it speaks to the *relative* and not the *absolute* sizes of the two coefficients (though here with similar mean values for ρ_s and ρ_i and similar distributions, absolute and relative variation are very close).

For a variable to have a high intracluster correlation coefficient there must be relative spatial concentration of the characteristic in the population. This means that individuals (or households) with a particular level or score on this variable choose to, or are forced to, or simply tend to, live near other individuals (households) that are similar to them. Consider the kinds of characteristics that we know are socio-economically clustered -- income, political persuasion, for example. The WFS has produced a classification of variables according to the median values of the clustering effect (see Verma, Scott, and O'Muircheartaigh, 1980). From that analysis it can be seen (in particular in table 4) that variables related to contraceptive knowledge and use are generally subject to much higher clustering effects than socio-demographic variables such as age and marital status.

The more clustered a variable is, the more likely it is to be a marker of the community. [I am here using the primary sampling unit (the cluster) as a proxy for community; the primary sampling units in the BHPS are relatively small; in WFS the cluster was usually either a village, a block, or in some cases a census enumeration area]. The mere fact of being relatively clustered within the PSU might make it more likely that the characteristic would be sensitive to social desirability effects, self-presentation effects, or status effects, especially when the interrogator is an outsider, as the interviewer typically would be. More importantly different interviewers could engender different levels of effects on the response, perhaps depending on the social distance between the interviewer and the respondent. In other words, this may mean that when these topics are broached by an outsider (the interviewer) the interviewer's role performance (and even the interviewer's extra-role characteristics) are (more) likely to influence the responses.

For the WFS data considered in chapter 4 corroborative evidence can be adduced to support the interpretation of, and belief in, the correlation between clustering effect and interviewer effect. The variables that have the highest values of ρ_i also tend to have high values of ρ_s. Not only that, but the variables with low values of ρ_i do not generally have high values of ρ_s.

These results raise the question again of the meaning of error, and our approach to its interpretation, in sample surveys. In computing sampling error, we consider a large value of

ρ to be a nuisance, as indeed it is if we are concerned only with the variance of the mean of a variable. We could however look at the value of ρ as information about the distribution of the variable in the population, and thus as additional information about the characteristic in which we are interested. In particular if our objective is to understand the distribution of the characteristic rather than merely to estimate its mean, then the value of ρ is important. The clusters we use in designing the sample are based on a real aspect of the population, its distribution and relative density, and sometimes also its social organization. To disregard this information completely is probably unwise.

There is a parallel in our treatment of the interviewer effect. If we find a pattern in the way in which interviewers provoke responses, or distortion of responses, then this pattern in itself contains information about the characteristic we are examining, and should not be treated merely as an undesirable factor that increases the variance of our estimators. This does not of course mean that it is not such a factor, simply that the phenomenon contains information additional to its impact on the variance.

The survey sampling perspective may be instrumental in keeping us from this broader perspective. By concentrating as it does on the estimation of the mean of a single variable at a time every influence is interpreted in terms of its impact on the variance of our estimate of this mean, narrowly interpreted. If instead we borrow the modelling perspective from other branches of statistics a different light is shed on the issue. From a modelling perspective any influence that causes the observations to depart from *iid* should be represented by one or more terms in the model. In this context, rather than treat the clustering effect as a variance the approach would be to include a term (or terms) representing its effect in whatever model was being proposed for the phenomenon being investigated. In the absence of an explicit integrated model, when only a mean is being estimated, the (standard statistical analysis of variance) approach is a variance components analysis. I discuss these issues further below.

11.2 Batteries of Items

The case of batteries of attitude items provides a vivid example of some of these principles

and approaches; not only can we look at the magnitude of effects, but we can contrast the impact on single items and combinations of items, and speculate about information to be extracted from the response errors.

It is common in survey research to have batteries of questions that are assumed on substantive grounds to be interrelated (this assumption is usually tested empirically at the pretest stage). Attitude scales are a prime example. In chapter 5 we consider the impact of the interviewer on (a) a set of attitude scales in a job satisfaction survey and (b) different sets of attitude items in a survey of impact of aircraft noise on residents of neighbourhoods close to an airport.

Operationally the most important finding was that the effect of interviewers on the variance of the battery score could not be predicted from the univariate effect of interviewers on the individual items. This also serves to concentrate attention on the use that is to be made on the data rather than on the traditional emphasis in survey research on means and totals for single variables. In the case of the airport noise data in particular the impact of the intrainterviewer correlated variance was much greater for the General Health Questionnaire (GHQ) scale than it was for the annoyance scale, even though the effect on the individual items in the GHQ was substantially less than the individual effects for annoyance; the GHQ is a widely used standard instrument for assessing psychiatric disturbance.

This difference in impact raises the question of how this intrainterviewer effect is generated. A simple theoretical exploration of the problem indicated that the crucial characteristic in determining the effect on a composite indicator would be the set of intercorrelations among the interviewer effects and the strength of these correlations relative to the strength of the intercorrelations among the items themselves (having removed or avoided intrainterviewer effects). The available empirical data showed examples of all possible outcomes.

As the ANOVA model expresses the basic component of interviewer effect as an additive relative interviewer bias, it seems natural to examine the set of these effects directly. Using an ANOVA model we can estimate for each item for each interviewer the magnitude of this component. This generates a matrix of effects, with one effect for each interviewer for each

item. This set of effects can be analyzed directly; of particular interest is the underlying dimensionality of the matrix. A principal component analysis is one way to investigate this. In each of the cases analyzed, we found a coherent structure to the effects. In none of them was there however a single dimension that dominated all the items. Had there been, that scale would be the most likely to show a greater impact for interviewer effect on the scale as a whole than on the items that make up the scale.

I argue that the pattern of interviewer effects (the elementary components designated by α_i) provides information about the substantive content of the attitude scale. A high intracluster correlation coefficient gives us information about the structure of the population; in a modelling context estimates of the individual cluster effects would tell us about how these clusters varied. The pattern of these cluster effects would contain information about the nature of the variation among clusters and would provide insights into the composition and determinants of the survey variable. In an analogous way, a high intrainterviewer correlation coefficient tells us about the sensitivity of the variable being examined to the social context of the data collection. The pattern of individual interviewer effects across items contains information about the way in which items have a common sensitivity to the characteristics of the data collection process. On particular if a subgroup of items has a common response to a set of interviewers, and this common response is different from the response to another set of items, this in itself potentially contains information on the subject matter being measured. For example if for one subset of items interviewers A,C, and D tended to induce an increase in level of agreement (compared to the effect of interviewers E,F, and G) and for a second subset of items the differentiation was between the effects of interviewers A and D with the other interviewers somewhere in between, this would indicate that there was some aspect of the subsets of items that generated this differential response. A second scenario of interest would be the case where the differentiation was among the same sets of interviewers but the direction of the net interviewer biases was reversed. From the theoretical analysis this would lead to an extremely efficient scale as these interviewer effects would tend to cancel each other out, thereby reducing substantially the total variance of the scale scores.

All these results point towards the desirability of broadening the consideration of interviewer

effect in two directions: first, the consideration of sets of items as a whole; and second, the potential of using the pattern of disturbances introduced into the data by the measurement process to understand both the measurement process and the subject matter more thoroughly.

11.3 Response Variability, the Respondent, and Social Relationships

At this stage it may be worth taking a step back and considering again the component underlying all the analysis of the response/measurement error in the 2x2 table of the components of the total variance. This is the simple response variance (SRV). Though we have so far considered this to be an irreducible element in the analysis, there is no reason why we cannot conceptualize the SRV as being itself composed of more elementary levels of measurement disturbances (as indeed it is described, though not elaborated, above). It is simply the difficulty of obtaining data to assess the elements of SRV that generally prevents us from doing so. The WFS data provide two traditional reinterview data sets with self-reports only. These permit some assessment of response variance for different subclasses of respondents [section 6.4 of chapter 6] and, by using the time interval between interview and reinterview, a crude estimation of between-trial correlation [chapter 8].

It is difficult in practice to separate the factors affecting the SRV from the factors affecting the estimation of the SRV (see figure 8.2 in chapter 8 for a schematic presentation). The very large scale data set from the US Current Population Survey (CPS) turns out to provide a vehicle to do so and to carry out an assessment of some of the components of SRV.

From two years' data (1982-84) there are some 120,000 reinterviews (re-enumeration in the terms of our classification of response variance designs) of individuals. The data set provides material on the impact of different influences (ranging from the target variables themselves to the circumstances of data collection) on the magnitude of the SRV. In one way this is simply a refinement (or a re-orientation) of the notions of the ANOVA model, where the α is the separated impact of the interviewer, but is also a component in the simple total variance; each influence has an α-like impact on the individual observations. [The correlated variance is the effect of any of these influences when it is affecting in the same way a set of respondents

or observations.]

The complex structure of the data collection process permits some analyses that would not be possible for a survey in which only self-reports were permitted and where the two interviews were independent. In particular it makes it possible to obtain a direct estimate of between-trial correlation of the measurement errors, to carry out a comparison of self and proxy reports, to assess some aspects of communication in the household, and to examine the quality of reports for respondents (and subjects) with different demographic characteristics. It also makes it possible to speculate (with empirical support) about both the training and attitudes of the interviewers and to gain some insight into the degree to which husbands and wives in the US pay attention to each others' utterances.

Between-trial correlation has always been an obstacle to the estimation of SRV. If the period between observations is too short, the responses are likely to be affected by the respondent's and/or the interviewer's recollection of the initial occasion. On the other hand if the period is too long, the risk of a real change in status is high, as is the difficulty of tracing the respondent. The WFS data provided some evidence of the effect of lengthening the time period. The CPS data provided two different ways of estimating the correlation directly. It turns out that without a more highly specified model the data provide too much information to choose between the two (the model is over-identified); both however indicated that the standard estimate of SRV underestimates the variance by more than 30%.

By considering in broadly the same way the comparison of the different combinations of respondents in the original interviews and the reinterviews it is possible to gain some insight into the assumptions and behavior of respondents and interviewers (and of the nature of communication in American families). A finding of some interest (section 8.5 of chapter 8) was that these data suggest that there is an asymmetry in communication in households. It would appear that wives are more likely to pay attention to (be influenced by) what their husbands say about being interviewed than husbands are to pay attention to (be influenced by) what their wives say. Furthermore it appeared that interviewers in general tended to give more credence to husbands' reports about their wives' employment status than to wives'

reports about their husbands' employment status. If true, these are important empirical statements about US society.

The other major set of results in chapter 8 relates to proxy respondents. When self-reports and proxy reports were compared across the full set of data the expected result was obtained: self-reports showed lower SRV than proxy reports. When the comparison was extended to particular subclasses of respondents - age and relationship to head of household - this finding collapsed: for each of the subclasses the direction of the difference was reversed - proxy reports had higher reliability than self-reports. This result challenges the conventional wisdom of survey design and execution.

The result is also counter-intuitive in that it is difficult to see how if every subclass is affected in the same way, the whole sample must not necessarily follow suit. The explanation of this latter point lies in the configuration of the respondents. The distribution of individuals across subcategories is very different for self-reports than for proxy-reports. For self-reports almost all the respondents are heads of household and wives of heads of household; for proxy-reports about 40 percent of the cases belong to the category of "other relatives". The latter group has much higher *gdr* for both self and proxy-reports than any other category, and dominates the *gdr* for the proxy-report category. This is what leads to the erroneous impression, given by the overall comparison, that proxy-reports are less reliable than self-reports. The apparent contradiction between the marginal (overall) contrast and the detailed subclass contrasts is an example of Simpson's Paradox (Simpson, 1951). The important point to bear in mind is that the subclass contrasts represent a more controlled comparison of self-reports and proxy-reports. The apparent superiority of self-reports suggested by the overall comparison is merely an artifact of the relationship between the frequency distribution of response by type of subject and the different levels of reliability obtained for different types of subject.

There remains the disturbing conclusion that proxy-reports are superior to self-reports; this contradicts the conventional wisdom that the best information comes from self-reports (see, for instance, Sudman and Bradburn, 1974). It is not suggested that simple response variance is a measure that describes all, or even the most important, aspects of quality. It is possible

that simple response variance may be lower but that biases, for instance, may be greater. There is no evidence in the data here to suggest that this is the case, but it is certainly a possibility that must be borne in mind. There is some other evidence emerging from work on the survey of Income and Program Participation (SIPP) and other major US studies that self-reports for factual data may not be greatly superior to proxy reports.

Finally, the CPS data provide evidence on the quality of information (in terms of reliability) provided by different socio-demographic categories of respondents; the WFS data also provide results on this topic. Unfortunately the same ancillary variables were not available for both surveys. The WFS data show that the less educated and the less urban subpopulations have higher simple response variance. The CPS data show that heads of household and their wives have lower response variance than others in the household. For CPS we get more reliable answers from older respondents, for WFS we get less reliable responses from older respondents. This apparent contradiction however may be explained by the nature of the questions, and the explanation may provide a corrective against simplistic conclusions.

In the WFS the questions (to women aged 15-49) referred to the dating of birth and marital histories. Thus for older women the dates being sought were farther in the past (more distant from the date of interview) making the cognitive task more difficult for older respondents. In the CPS the questions referred to employment status; with increasing age, the stability of employment status increases (and for older respondents who are retired becomes completely stable) and the incidence of ambiguous employment situations decreases; thus the cognitive task typically would be more difficult for younger respondents. It is worth bearing in mind that the relative difficulty of different tasks varies across subgroups of the population and that therefore not all patterns across demographic subclasses will be stable.

The effect of motivation on the performance of the tasks will introduce additional complexity. Though the results here do not address motivation directly, the WFS data (section 7.1 of chapter 7) produce a measure of respondents' *Cooperation*, as judged by the interviewers. There is an unambiguous negative correlation between the perceived level of *Cooperation* and

the magnitude of the response deviations in both Lesotho and Peru, confirming that the level of cognitive difficulty is not the only factor affecting response quality.

These analyses (in chapters 6, 7, and 8) are restricted in that they do not consider the joint effects of the factors examined. In order to do this it is necessary to consider explicit modelling of the whole process. The purpose of the analyses described in chapter 9 is to do this, and to provide a framework within which the design and measurement issues can be disentangled. In chapter 10, the substantive and methodological analyses are combined in a single structure.

11.4 The Modelling Perspective

All analysis is modelling; even the calculation of a mean implies that there is enough homogeneity in a population to make its central tendency informative. All analyses of response variance similarly imply some assumptions about the process that generates response errors; the model in chapter 2 section 2.3 is just one possibility. But despite this inescapable model dependence, there has been a tendency among survey researchers to eschew models as much as possible.

This dissertation charts a journey, that began in the early 1970s, away from this tendency. The seeds of the departure can be found in the examination of the multivariate structure of interviewer effect (O'Muircheartaigh, 1974 and section 5.2 of chapter 5). That work takes a two-stage approach to interviewer effect, starting with a multivariate analysis of variance and following it with a principal component analysis of the estimated individual interviewer effects. Even then I realised that it would be better to combine the stages in a single operation, but lacked the methodology to do so. The analyses also demonstrate that it is possible in principle to derive insights into the subject matter of the survey from the pattern of distortions in the data.

In some ways the material on batteries of items in a scale (O'Muircheartaigh, 1976 and section 5.1 of chapter 5) also contains elements of later analyses. An attitude scale is a set of

items (variables) where we have a strong model assumption that the items belong together; the individual items have no existence except as components of the scale. Consequently all analyses of the items -- including analyses of response errors -- must recognize this reality. The logical conclusion of this line of argument is that analyses of response variance should also be directed at the results of analyses and not at variables individually.

Even the standard analyses of the sampling statistician contain the germs of modelling. An important aspect of the computation of design effects is in the impact on subclasses and the comparison of subclasses; earlier work on sampling errors in the WFS emphasized these issues. Verma, Scott, and O'Muircheartaigh (1980), O'Muircheartaigh (1984a, 1984b, and 1986), section 6.4 in chapter 6, and section 8.4 in chapter 8 all indicate that results for the whole population do not necessarily follow through to particular subpopulations. To a modeller this implies that the characteristics by which the results vary should be incorporated specifically into a modelling framework.

It is one thing to recognize the need for modelling, it is another to find the right way to conceptualize the issues so that models help to clarify them. I have so far followed two different paths, one inwardly directed at the survey process, the other outwardly directed at the survey results.

The first approach focusses on the measurement process itself and models the factors that determine its outcome. In O'Muircheartaigh (1991) and chapter 9 we model explicitly the determinants of the SRV for binary variables. The results are encouraging as they allow simultaneous consideration of multiple factors. In particular they make it possible to model both the quality of the response process itself (the SRV) and the data collection process that can contaminate the estimates of the effect of the response process (memory effects and communication effects for example). This approach clarifies many of the confusions that arose from the univariate analyses in chapter 8.

The second approach assumes that the purpose of collecting the survey data is to carry out complex analyses of the data, and in particular to model relationships among the variables.

It takes as its aim the incorporation of the response errors directly into these analyses. It is the natural outgrowth of the preceding work.

There are three stages in the progression of this work. In O'Muircheartaigh and Wiggins (1981) and section 10.1 of chapter 10, the interviewer is included as an explicit factor in the models of the relationships among factors related to noise annoyance. The results are provocative in that they demonstrate that even with variables that are subject to substantial interviewer effects there may be no impact on analyses that include these variables. The analysis is restricted in usefulness however as, with a larger number of interviewers and without complete interpenetration, this kind of model would become either unwieldy or impossible.

Section 10.2 applies a more general model to the data treated in section 10.1. Here a general hierarchical linear model is used that treats the interviewer as a grouping factor in the data and includes it explicitly in the model. In section 10.3 similar models are applied to the BHPS data but with the inclusion of CSV (correlated sampling variance). The results confirm the feasibility of integrating both the sample design and the fieldwork design in substantive modelling of effects.

11.5 Conclusion

We can see from the preceding material that there are really three separate strands of argument here, each associated with a particular approach to the analysis. First there is the descriptive (simple diagnostic) orientation of establishing the circumstances under which (or if) response variability, interviewer effects and clustering effects occur, the associated issue of how they should be accommodated in analysis - primarily estimating the impact on the variance of univariate statistics - and an assessment of their likely order of magnitude. Second, there is the model-assisted orientation which attempts to decompose the effects into their constituent parts. There are two sub-strands within this approach: one is to incorporate the correlating source (cluster or interviewer for example) as a term or terms in other models that we are estimating so that the effect is incorporated into the estimation of these models;

the other is to model the response error itself -- in doing this we are trying to decompose it into its constituent parts. Third, and most radical, is to view error as information. By conceptualizing the process that generated the errors as a substantive process rather than as a set of nuisance effects we can extract from the results of the process information about both the process and the subject matter. Any particular piece of analysis may include any combination of these three approaches.

I would like to feel that the broad tenor of this dissertation has been to emphasise that survey methodology is just one part of the survey process and that the sample survey and its results should be seen in their broader context. In the framework presented in section 1.2, this means that any assessment of social data should take into account the positioning of the inquiry in relation to the three dimensions of *representation*, *randomization*, and *realism*, as well as considering the more traditional elements of interviewer, respondent, and task.

Bowley expressed, in his 1915 book on measurement, a suitably broad perspective of the issues:

"The main task ... is to discover exactly what is the critical thing to examine, and to devise the most perfect machinery for examining it with the minimum of effort ... . In conclusion, we ought to realise that measurement is a means to an end; it is only the childish mind that delights in numbers for their own sake. On the one side, measurement should result in accurate and comprehensible description, that makes possible the visualization of complex phenomena; on the other, it is necessary to the practical reformer..."

This inclusion in the definition of measurement of both the substantive decisions about what to examine and the technical issues of how to examine it is a useful reminder that there is a danger that those directly involved in the measurement process can at times become too restricted in their vision by the technical fascination of their own discipline. We need to step outside those confines and acknowledge that others may have different requirements from what we think of as *our* data. Survey methodologists should become more *externalist* and less *internalist* (see Converse, 1986) and accept that they should respect others' criteria as

well as their own.

This dissertation is not the final word on response variance in sample surveys; it is simply a stage along the road. It would be desirable to knit together the elements of the arguments into a seamless whole, and in particular to find a way of translating error, or the understanding of error, into an additional aspect of the information derived from a survey. Attempting to find that understanding should provide adequate work for some time to come.

Aitkin, M., Anderson, D., and Hinde, J. (1981) Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society, A*, Vol. 144, pp. 419-461

Alam, I. and Cleland, J. (1981) Illustrative analysis: recent fertility trends in Sri Lanka. *WFS Scientific Report* no 25. The Hague: International Statistical Institute.

Anderson, D., and Aitkin, M. (1985). Variance component models with binary response. *Journal of the Royal Statistical Society, B*, Vol. 47, pp. 203-210.

Anderson, R.L. and Bancroft, T.A. (1952). *Statistical Theory in Research* . New York: McGraw-Hill.

Bailar, B.A. (1968). Recent research in reinterview procedures. *Journal of the American Statistical Association*, 63, 41-63.

Bailar, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-30.

Bailar, B.A. and Rothwell, N.D. (1984). Measuring employment and unemployment. In *Surveying Subjective Phenomena* (C.F. Turner and E.A. Martin (eds.))Vol.2, pp. 129-142. New York: Russell Sage Foundation.

Bailey, L., Moore, T.F., and Bailar, B.A. (1978). An interviewer variance study for the eight impact cities of the National Crime Survey cities sample. *Journal of the American Statistical Association*, **73**, 16-23.

Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., Sudman, S. (eds.) (1991) *Measurement Errors in Surveys*. New York: John Wiley and Sons.

Bartholomew, D.J. (1987). *Latent Variable Models and Factor Analysis*. London: Griffin.

Bingham, W.V., and Moore, B.V. (1934). *How to Interview*, revised edition. New York: Harper.

Blalock, H.M., and Blalock, A.B., (eds.) (1968). *Methodology in Social Research.* New York: McGraw-Hill.

Bogardus, E.S. (1925). Measuring social distances. *Journal of Applied Sociology*, 9, 299-308.

Booker, H.S. and David, S.T. (1952). Differences in results obtained by experienced and inexperienced interviewers. *Journal of the Royal Statistical Society, A*, 115, 232-257.

Booth, C., (ed.) (1889). *Labour and the Life of the People of London.* London: Macmillan

Booth, H. and Alam, I. (1984). The Pakistan Fertility Survey: The quality of the data. *WFS Monograph on Pakistan.* The Hague: International Statistical Institute.

Bowley, A.L. (1906). Presidential Address to the Economic Section of the British Association for the Advancement of Science. *Journal of the Royal Statistical Society*, 69, 540-558.

Bowley, A.L. (1923). *The Nature and Purpose of the Measurement of Social Phenomena,* 2nd edition. London: P.S. King and Son, Ltd.

Bowley, A.L. (1926). Measurement of the precision attained in sampling. *Bulletin of the International Statistical Institute*, 22, 1, 1-62 of special annex following p. 451.

Brass, W. (1978). Screening procedures for detecting errors in maternity history data. *Unpublished manuscript.*

Bryk, A.S., and Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Newbury Park: Sage.

Bryk, A.S., Raudenbush, S.W., Congdon, R. and Seltzer, M. (1986). *An Introduction to HLM: Computer Program and User's guide.* Chicago: University of Chicago, Department of Education.

Campanelli, P C and O'Muircheartaigh, C (1999). Interviewers, interviewer continuity, and panel study nonresponse. *Quality and Quantity*, 33, 59-76.

Campbell, D.T., and Stanley, J.C. (1963). *Experimental and Quasi-experimental Designs for Research*. Chicago: Rand McNally.

Cannell, C.F. and Fowler, F.J. (1963). A comparison of self-enumerative procedure and a personal interview: A validity study. *Public Opinion Quarterly*, 27, 250-264.

Cannell, C.F., Groves, R. and Miller, P.V. (1981). The effects of mode of data collection on health survey data. *Proceedings of the Social Statistics Section*, American Statistical Association, 1-6.

Cannell, C.F., Kalton, G., Oksenberg, L., Bischoping, K., and Fowler, F. (1989). *New Techniques for Pretesting Survey Questions* (unpublished). Final Report for grant HS05616, National Center for Health Services Research, Ann Arbor: University of Michigan.

Cannell, C.F., Lawson, S.A., and Hausser, D.L. (1975). *A Technique for Evaluating Interviewer Performance*. Ann Arbor: Institute for Social Research, The University of Michigan.

Cannell, C.F., Marquis, K.H. and Laurent, A. (1970). *A Summary of Research Studies in Interviewing Methodology*. Ann Arbor: Survey Research Centre.

Cannell, C.F., Miller, P.V., and Oksenberg, L.F. (1981). Research on interviewing techniques. in S. Leinhardt (ed.) *Sociological Methodology*, pp. 389-437. San Francisco: Jossey-Bass.

Cantril, H., and Associates (1944). *Gauging Public Opinion.* Princeton: Princeton University Press.

Cartwright, A. (1957). A method of obtaining information from different informants on a family morbidity inquiry. *Applied Statistics*, 6, 18-25.

Chai, J. (1971). Correlated measurement errors and the least squares estimator of the regression coefficient. *Journal of the American Statistical Association*, 66, 478-483.

Chidambaram, V.C., Cleland, J.G. and Verma, V.K. (1980). Some issues of survey methodology and data quality: the WFS experience. Paper prepared for the PAA Meeting, Denver.

Choi, I.C. and Comstock, G.W. (1975). Interviewer effect on responses to questionnaire relating to mood. *American Journal of Epidemiology*, 101, 84-92.

Chua, T.C. and Fuller, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, **82**, 46-51.

Cicchetti, D.V. (1972). A new measure of agreement between rank ordered variables. *Proceedings, 80th Annual Convention, APA*: 17-18.

Cicchetti, D.V. and Allison, T. (1973). Assessing the reliability of scoring EEG sleep records: an improved method. *Proceedings and Journal of the Electro-Physiological Technologists' Association 20*: 92-102.

Cicourel, A., *Method and Measurement in Sociology*, New York: Free Press, 1964.

Cochrane, A. *et al* (1951). Observer errors in taking medical histories. The Lancet, 1007-1009.

Cochran, W.G. (1953). *Sampling Techniques*, New York: John Wiley and Sons; 2nd edition 1963; 3rd edition 1977.

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

Cochran, W.G. (1970). Some effects of errors of measurement on multiple correlation. *Journal of the American Statistical Association*, 65, 22-34.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. and Psych. Meas. 20*: 92-102.

Cohen, J. (1968). Weighted kappa, nominal scale agreement with provision for scaled disagreement or partial credit. *Psych. Bull 70*: 213-22.

Collins, M. (1980). Interviewer variability: A review of the problem. *Journal of the Market Research Society*, 22.

Collins, M. and Butcher, B. (1982) Interviewer and clustering effects in an attitude survey. *Journal of the Market Research Society*, **25** (1), 39-58.

Converse, J.M. (1986). *Survey Research in the United States: Roots and Emergence 1890-1960*. Berkeley: University of California Press.

Deming, W.E. (1960). *Some Theory of Sampling*. New York: Wiley.

Dempster, A.P., Laird, N.M. and Tsutakawa, R.K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association*, 76, 341-353.

Dijkstra, W., and van der Zouwen, J., (eds.) (1982). *Response Behaviour in the Survey-Interview*. London: Academic Press.

Dohrenwend, B.P. (1966). Social status and psychiatric disorder: An issue of substance and an issue of method. *American Sociological Review*, 31, 14-34.

DuBois, W.E.B. (1899). *The Philadelphia Negro: A Social Study*. Philadelphia: Ginn.

Durbin, J. and Stuart, A. (1951). Differences in response rates of experienced and inexperienced interviewers. *Journal of the Royal Statistical Society, A*, 114, 163-205.

Eckler, A.R., and Pritzger, L. (1951). Measuring the accuracy of enumerative surveys. *Bulletin of ISI*, 33/4. 7-24.

Fairbairn, A.S. *et al* (1959). Variability in answers to a questionnaire on respiratory symptoms. *British Journal of Preventative Soc. Medicine*, 17, 175-193.

Feather, J. (1973). *A Study of Interviewer Variance*. Saskatoon, Canada: Department of Social and Preventive Medicine, University of Saskatchewan.

Fellegi, I. P. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, **59**, 1016-1041.

Fellegi, I.P. (1974). An improved method of estimating the correlated response variance. *Journal of the American Statistical Association*, **69**, 496-501.

Fellegi, I.P. (1980). Approximate tests of independence and goodness of fit based on stratified multistage samples. *Journal of the American Statistical Association, 75*, 261-8.

Fleiss, J.L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. and Psych. Meas. 33*: 613-19.

Florez, C.E. and Goldman, N. (1980). An analysis of nuptiality data in the Colombia National Fertility Survey. *WFS Scientific Report no. 11.* The Hague: International Statistical Institute.

Freeman, J. and Butler, E.W. (1976) Some sources of interviewer variance in surveys. *Public Opinion Quarterly*, **40**, 79-91.

Gales, K.E., and Kendall, M.G. (1957) An inquiry concerning interviewer variability. *Journal of the Royal Statistical Society*, Series A, **120**, 121-147.

Galton, F. (1883). *Inquiries into the Human Faculty* quoted in Ruckmick (1930) *op. cit.*

Gates, G.S., and Rissland, L.Q. (1923). The effect of encouragement and of discouragement upon performance. *Journal of Educational Psychology*, 14, 21-26.

Giesbrecht, F.G. and Burrows, P.M. (1978). Estimating variance components in hierarchical structures using MINQUE and restricted maximum likelihood. *Comm. Statist. A*, 7, 891-904.

Gillin, J.L. (1915). The social survey and its further development. *Journal of the American Statistical Association*, 14, 603-610.

Godambe, V.P.(1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17, 269-278.

Goldberg, D.P. (1979). *The detection of psychiatric illness by means of a questionnaire.* London: Oxford University Press.

Goldman, N., Coale, A.J. and Weinstein, M. (1979). The quality of data in the Nepal Fertility Survey. *WFS Scientific Reports no 6.* The Hague: International Statistical Institute.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Griffin.

Goldstein, H. (1991). Nonlinear multilevel models. *Biometrika*, 78, 45-52.

Goldstein, H. (1995) *Multilevel Statistical Models,* second edition. London: Edward Arnold; New York: Halstead Press.

Goodman, L. (1973). Causal analysis of data from panel studies and other kinds of surveys. *American Journal of Sociology*, 82, 6, 1289-1317.

Gove, W.R. and Geerken, M. (1976). Response bias in surveys of mental health: An epidemiological investigation. *American Journal of Sociology*, 82, 6, 1289-1317.

Gray, P.G. (1955). The memory factor in social surveys. *Journal of the American Statistical Association*, 50, 344-363.

Gray, P.G. (1956). Examples of interviewer variability taken from two sample surveys. *Applied Statistics*, **5**, 73-85.

Groves, R.M. (1989). *Survey Errors and Survey Costs*. New York: John Wiley.

Groves, R.M. and Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, **50**, 251-256.

Guzman, J.M. (1980). Evaluation of the Dominican Republic National Fertility Survey 1975. *WFS Scientific Reports no 14.* The Hague: International Statistical Institute.

Hansen, M.H., Hurwitz, W.N. and Bershad, M.A. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, **38**(2), 359-374.

Hansen, M.H., Hurwitz, W. and Pritzger, L. (1964). The estimation and interpretation

of gross differences and the simple response variance. *Contributions to Statistics*. Calcutta: Pergamon.

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

Hanson, R.H. and Marks, E.S. (1958). Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, **53**, 635-55.

Hartley, H.O. and Rao, J.N.K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93-108.

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

Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61, 383-385.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320-340.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226, 252.

Hill, D. (1987). Response errors in labour surveys: comparisons of self and proxy reports in the SIPP. *Proceedings of the Third Annual Research Conference*, U.S. Bureau of the Census, 299-319.

Hobcraft, J. (1980). Illustrative analysis: evaluating fertility levels and trends in Colombia. *WFS Scientific Reports no 15*. The Hague: International Statistical Institute.

Hochstim, J.R. (1967). A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association*, 62, 976-989.

Hovland, C.I., and Wonderlic, E.F. (1939). Prediction of industrial success from a standardized interview. *Journal of Applied Psychology*, 23, 537-546.

Hox, J.J., de Leeuw, E.D., and Kreft, I.G.G. (1991). The effect of interviewer andrespondent characteristics on the quality of survey data: a multilevel model. In *Measurement Errors in Surveys* (eds P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman). New York: John Wiley and Sons, Inc.

HYMAN, H.H. (1954). *Interviewing in Social Research*, Chicago: University of Chicago Press.

Jabine, T.B., Loftus, E., Straf, M.L., Tanur, J.M., and Tourangeau, R., (eds.) (1984). *Cognitive Aspects of Survey Methodology: Building an Bridge Between Disciplines.* Washington, D.C.: National Academic Press.

Jenrich, R.I. and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42, 805-820.

Joreskog, K.G. (1973). Analyzing psychological data by structural analysis of covariance matrices. In C. Atkinson, H. Krantz, R.D. Luce, and P. Suppes (eds.), *Contemporary Developments in Mathematical Psychology* pp. 1-56. San Francisco: Freeman.

Joreskog, K.G., and Sorbom, D. (1989). *LISREL 7: A Guide to the Program and Applications*. The Netherlands: SPSS International BV.

Kahn, R.L. and Cannell, C.F. (1957). *The Dynamics of Interviewing: Theory, Technique and Cases*. New York: Wiley.

Kalton, G. (1977). Practical methods for estimating survey sampling errors. *Bull. Int. Statist. Inst. 47/3.*

Kalton, G. (1979). Ultimate cluster sampling. *Journal of the Royal Statistical Society,* **142**(2), 210-222.

Katz, D. (1942). Do interviewers bias poll results? *Public Opinion Quarterly*, 6, 248-268.

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

Kemsley, W.F.F. (1960). Interviewer variability and a budget survey. *Journal of the Royal Statistical Society, A*, 128, 118-139.

Kiaer, A.N. (1897). *The Representative Method of Statistical Surveys*. Oslo: Kristiania.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

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

Kish, L.(1987). *Statistical Design for Research*, New York: John Wiley and Sons.

Kish, L. and Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society, Series B, 36*, 1-37.

Kish, L., Groves, R.M., and Krotki, K. (1976). Sampling errors for fertility surveys. *WFS Occasional Papers no 17.* The Hague: International Statistical Institute.

Kish, L. and Lansing, J. (1954). Response errors in estimating the value of homes. *Journal of the American Statistical Society*, 49, 520-538.

Klassen, D., Hornstra, R. and Anderson, P. (1975). The influence of social desirability on symptom and mood reporting in a community survey. *Journal of Consulting and Clinical Psychology*, 43, 4, 448-452.

Koch, G. (1969). The effect of non-sampling errors on measures of association in 2x2 contingency tables. *Journal of the American Statistical Society*, 64, 852-863.

Kruskal, W., and Mosteller, F. (1980). Representative sampling IV: the history of the concept in statistics. *International Statistical Review*, 48, 169-195.

Landis, J.R. and Koch, G. (1976). A review of statistical methods in the analysis of data arising from observer reliability studies. *Statistica Neerlandica 29*: 101-23 and 151-61.

Likert, R. (1932). *A Technique for the Measurement of Attitudes* (Archives of Psychology, no. 140). New York: Columbia University Press.

Likert, R. (1969). *CORE questionnaire*. Ann Arbor: Institute for Social Research, University of Michigan.

Lindstrom, M.J. and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83, 1014-1022.

Longford, N.T. (1986). Normal variance component modelling and exponential family extensions. *Proceedings of the section on Statistical Computing of the American Statistical Association.* 137-142. Alexandria, VA: American Statistical Association.

Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74, 817-827.

Longford, N.T. (1988). *VARCL Manual.* Princeton, New Jersey: Educational Testing

Service.

Longford, N.T. (1993). *Random Coefficient Models*. Oxford: Clarendon Press, 1993.

Lynn, P. and Lievesley, D. (1991). *Drawing General Population Samples in Great Britain.* London: Social and Community Planning Research.

Mahalanobis, P.C. (1944). On large scale sample surveys. *Roy. Soc. Phil. Trans.*, B, 231, 329-451.

Mahalanobis, P.C. (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, **109**, 325-70.

Mathiowetz, N. and Groves, R.M. (1983). The effects of respondent rules on health survey reports. In C.F. Cannell et al., *An Experimental Comparison of Telephone and Personal Health Surveys*, 83-102. Washington, D.C.: National Center for Health Statistics.

Mathiowetz, N. and Groves, R.M. (1985). The effects of respondent rules on health survey reports. *American Journal of Public Health*, **75**, 639-644.

McKennel, A. (1973). Psychological factors in aircraft noise annoyance. *Proceedings of the International Congress on noise as a public health problem (Dubrovnik)*, 627-644. Washington, D.C.: US Environmental Protection Agency Publication 550/973 008.

McNemar, Q. (1962). *Psychological Statistics*. New York: Wiley.

Moore, J.C. (1985). Self/proxy response status and survey response quality. *Unpublished manuscript*. Washington D.C.: U.S. Bureau of the Census,

Moore, J.C. and Marquis, K.H. (1988). Using administrative record data to describe SIPP response errors. Paper presented to the *Section of Survey Research Methods at the*

Meetings of the American Statistical Association, August 1988.

Moss, L., and Goldstein, H., (eds.) (1979). *The Recall Method in Social Surveys*. London: Institute of Education, University of London.

Muscio, B. (1917). The influence of the form of a question. *The British Journal of Psychology*, 8, 351-389.

Nelder, J. and Baker, B. (1981). *GLIM3 Manual.* Oxford: NAG Publications.

Neter, J. and Waksberg, J. (1964). A study of response errors in expenditures data from household interviews. *Journal of the American Statistical Society*, 59, 18-55.

Neter, J. and Waksberg, J. (1965). *Response Errors in Collection of Expenditure Data by Household Interviewers: An Experimental Study.* Washington D.C.: U.S. Government Printing Office.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

O'Muircheartaigh, C.A. (1974). "The structure of interviewer effect", Appendix 2 in O'Muircheartaigh, Colm (1974) *Withdrawal Behavior in Irish Manufacturing Industry.* Report to the Human Sciences Committee. Dublin: Institute for Public Administration.

O'Muircheartaigh, C.A. (1976). Response errors in an attitudinal survey. *Quality and Quantity*, 10, 97-115.

O'Muircheartaigh, C.A. (1977). Response errors. In C.A. O'Muircheartaigh and payne, C. (Eds.), *The Analysis of Survey Data*, volume II, *Model Fitting*, 193-239. London: John Wiley and Sons.

O'Muircheartaigh, C.A. (1982). Methodology of the Response Errors Project. *WFS Scientific Reports no 28.* The Hague: International Statistical Institute.

O'Muircheartaigh, C.A. (1984a) The magnitude and pattern of response variance in the Peru Fertility Survey. *WFS Scientific Reports No. 45.* The Hague: International Statistical Institute.

O'Muircheartaigh, C A (1984b) The magnitude and pattern of response variance in the Lesotho Fertility Survey. *WFS Scientific Reports No. 70.* The Hague: International Statistical Institute.

O'Muircheartaigh, C.A. (1986). Correlates of reinterview response inconsistency in the Current Population Survey (CPS). *Proceedings of the Second Annual Research Conference*, 208-234. Washington, D.C.: U.S. Bureau of the Census.

O'Muircheartaigh, C. A. (1991). Simple response variance: estimation and determinants. In Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., Sudman, S. (eds.) *Measurement Errors in Surveys*, 551-574. New York: John Wiley and Sons.

O'Muircheartaigh, C. A. (1997). Measurement errors in surveys: a historical perspective. In L Lyberg, P Biemer, M Collins, E de Leeuw, C Dippo, N Schwarz, and D Trewin (eds) *Survey Measurement and Process Quality*, pp1-25. New York: John Wiley and Sons.

O'Muircheartaigh, C. A. (1999). CASM [Cognitive Aspects of Survey Methodology]: successes, failures, and potential. In M G Sirken, D J Herrmann, S Schechter, N Schwarz, J M Tanur, and R Tourangeau (eds), *Cognition and Survey Research*, 39-62. New York: John Wiley and Sons.

O'Muircheartaigh, C. A. and Campanelli, P. C. (1998). The relative impact of sampling and interviewer effect on survey variance. *Journal of the Royal Statistical Society, Series A,* 161, 1, 63-78.

O'Muircheartaigh, C. A. and Campanelli, P. C. (1999). A multilevel exploration of the role of interviewers in survey nonresponse. *Journal of the Royal Statistical Society, Series A*, 162, 3, 437-446.

O'Muircheartaigh, C.A., and Soon, T.W.(1981). The impact of sampling theory on survey sampling practice: a review. *Bulletin of the International Statistical Institute*, 49, 1, 465-493.

O'Muircheartaigh, C.A. and Wiggins, R.D. (1981). The impact of interviewer variability in an epidemiological study. *Psychological Medicine*, 11, 817-824.

Pannekoek, J. A mixed model for analyzing measurement errors for dichotomous variables. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds.) *Measurement Errors in Surveys*, 517-530. New York: John Wiley & Sons, Inc.

Parlin, C.C. (1915). *The Merchandising of Automobiles, An Address to Retailers*. Philadelphia: Curtis Publishing Co.

Parry, H.J. and Crossley, M.J. (1950). Validity of responses to survey questions. *Public Opinion Quarterly*, 14, 61-80.

Patrick, D. (1981). *The health and care of the physically disabled. The longitudinal disability survey. Phase 1 Report.* London: Department of Community Medicine, St. Thomas' Hospital Medical School.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545-554.

Phillips, D. and Clancy, K. (1972). Some effects of social desirability in survey studies. *American Journal of Sociology*, 77, 921-940.

Potter, J.E. (1975). *The Validity of Measuring Change in Fertility by Analyzing Birth Histories Obtained in Surveys*. Doctoral dissertation. Princeton NJ: Princeton University.

Rasbash, J., Woodhouse, G., Goldstein, H., Yang, M., Howarth, J., and Plewis, I. (1995). *MLn Software*. London: Multilevel Models Project, Institute of Education, University of London.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, 13, 85-116.

Raudenbush, S.W. and Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59, 1-17.

Rice, S.A. (1929). Contagious bias in the interview. *American Journal of Sociology*, 35, 420-423.

Rowntree, B.S. (1902). *Poverty: A Study of Town Life*. London: Longmans.

Ruckmick, C.A. (1930). The uses and abuses of the questionnaire procedure. *Journal of Applied Psychology*, 14, 32-41.

Schaeffer, N.C. (1991). Conversation with a purpose - or conversation? Interaction in the standardized interview. In P.P. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, 367-391. New York: John Wiley & Sons, Inc.

Scheuch, E.K. (1967). Das Interview in der Sozialforschung. In R. Konig (ed.), *Handbuch der empirischen Sozialforschung*, 1, 136-196. Stuttgart: F. Enke.

Schuman, H., and Presser, S. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic Press.

Schwarz, N., Strack, F., and Mai, H. (1991). Assimilation and contrast effects in part-whole question sequences: a conversational analysis. *Public Opinion Quarterly*, 55, 3-23.

Sheatsley, P.B. (1947-8). Some uses of interviewer-report forms. *Public Opinion Quarterly*, 11, 601-611.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, B*, 13, 238-241.

Smith, T M F (1976). The foundations of survey sampling: a review (with discussion). *Journal of the Royal Statistical Society, Series A*, 139, 183-204.

Somoza, J.L. (1980). Illustrative analysis: infant and child mortality in Colombia. *WFS Scientific Reports no 10.* The Hague: International Statistical Institute.

Stouffer, S., and Associates (1949). *The American Soldier, Vol.1, Adjustment During Army Life*. Princeton: Princeton University Press.

Stouffer, S., and Associates (1949). *The American Soldier, Vol.2, Combat and Aftermath.* Princeton: Princeton University Press.

Strenio, J.F., Weisburg, H.I. and Bryk, A.S. (1983). Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics*, 39, 71-86.

Suchman, L., and Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, 85, 232-241.

Sudman, S., and Bradburn, N.(1974). *Response Effects in Surveys: a Review and Synthesis*. Chicago: Aldine Publishing Co.

Sukhatme, P.V. (1953). *Sampling Theory of Surveys with Applications*. New Delhi: The

Indian Society of Agricultural Statistics.

Sukhatme, P.V. and Sukhatme, B.V. (1970). *Sampling Theory of Surveys with Applications*. Bombay: Asia Publishing House.

Tanur, J. (ed.) (1992). *Questions About Questions: Inquiries into the Cognitive Bases for Surveys*. New York: Russell Sage Foundation.

Tarnopolsky, A., Barker, S.M., Wiggins, R.D. and McLean, E.K. (1978). The effect of aircraft noise on the mental health of a community sample: A pilot study. *Psychological Medicine*, 8, 219-233.

Thomas, W.I. (1912). Race psychology: standpoint and questionnaire, with particular reference to the immigrant and the negro. *American Journal of Sociology*, 17, 725-775.

Thompson, R. and Meyer, K. (1986). Estimation of variance components: What is missing in the EM algorithm?. *Journal of Statist. Simul.*, 24, 215-230.

Thurstone, L.L. (1929). Theory of attitude measurement. *Psychological Bulletin*, 36, 222-241.

Tourangeau, R., and Rasinski, K.A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103, 299-314.

Trussell, J. (1980). Illustrative analysis: age at first marriage in Sri Lanka and Thailand. *WFS Scientific Reports no 13*. The Hague: International Statistical Institute.

U.S. BUREAU OF THE CENSUS (1968). *Evaluation and Research Program f the U.S. Censuses Population and Housing, 1960: Effects of Interviewers and Crew Leaders*. Washington D.C.: U.S. Government Printing Office.

Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey. *Journal of the Royal Statistical Society, A*, **143**(4), 431-473.

Waterton, J. and Lievesley, D. (1987). Attrition in a panel survey of attitudes. *Journal of Official Statistics*, **3**, 267-282.

Wechsler, J. (1940). Interviews and interviewers. *Public Opinion Quarterly*, 4, 258-260.

Weiss, C. (1968). Validity of welfare mothers' interview responses. *Public Opinion Quarterly*, 32, 622-633.

Wiggins, R.D. (1979). Sample design for West London Survey of aircraft noise. In A. Tarnopolsky and J. Morton Williams (Eds.), *Aircraft Noise and Psychiatric Morbidity*, 12-29. London: SCPR.

Wiggins, R.D. (1985). A replicated study of the impact of interviewer variability in a community survey of physically handicapped in an Inner London Borough. *Research Working Paper No. 24*. London: Polytechnic of Central London.

Wiggins, R.D., Longford, N., and O'Muircheartaigh, C.A. (1992). A variance components approach to interviewer effects. In A. Westlake, R. Banks, C. Payne, and T. Orchard (eds.) *Survey and Statistical Computing*. Amsterdam: North-Holland.

Williams, D.(1942). Basic instructions for interviewers. *Public Opinion Quarterly*, 6, 634-641.

Woodhouse, G. (1995). *A Guide to MLn for New Users*. London: Multilevel Models Project, Institute of Education, University of London.

Yates, F. (1949). *Sampling Techniques for Censuses and Surveys*. London: Griffin.

Yuker, H.E. *et al* (1970). The measurement of attitudes towards disabled persons, *Rehabilitation Series 3, Insurance Company of North America.* New York: Ina Mend Institute of Human Resources.

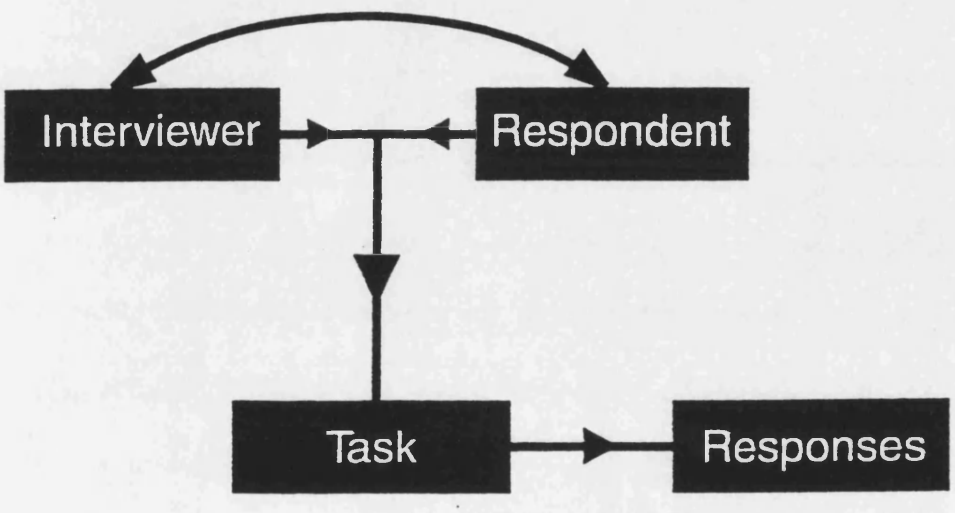


Figure 1.1 A model of the survey interview

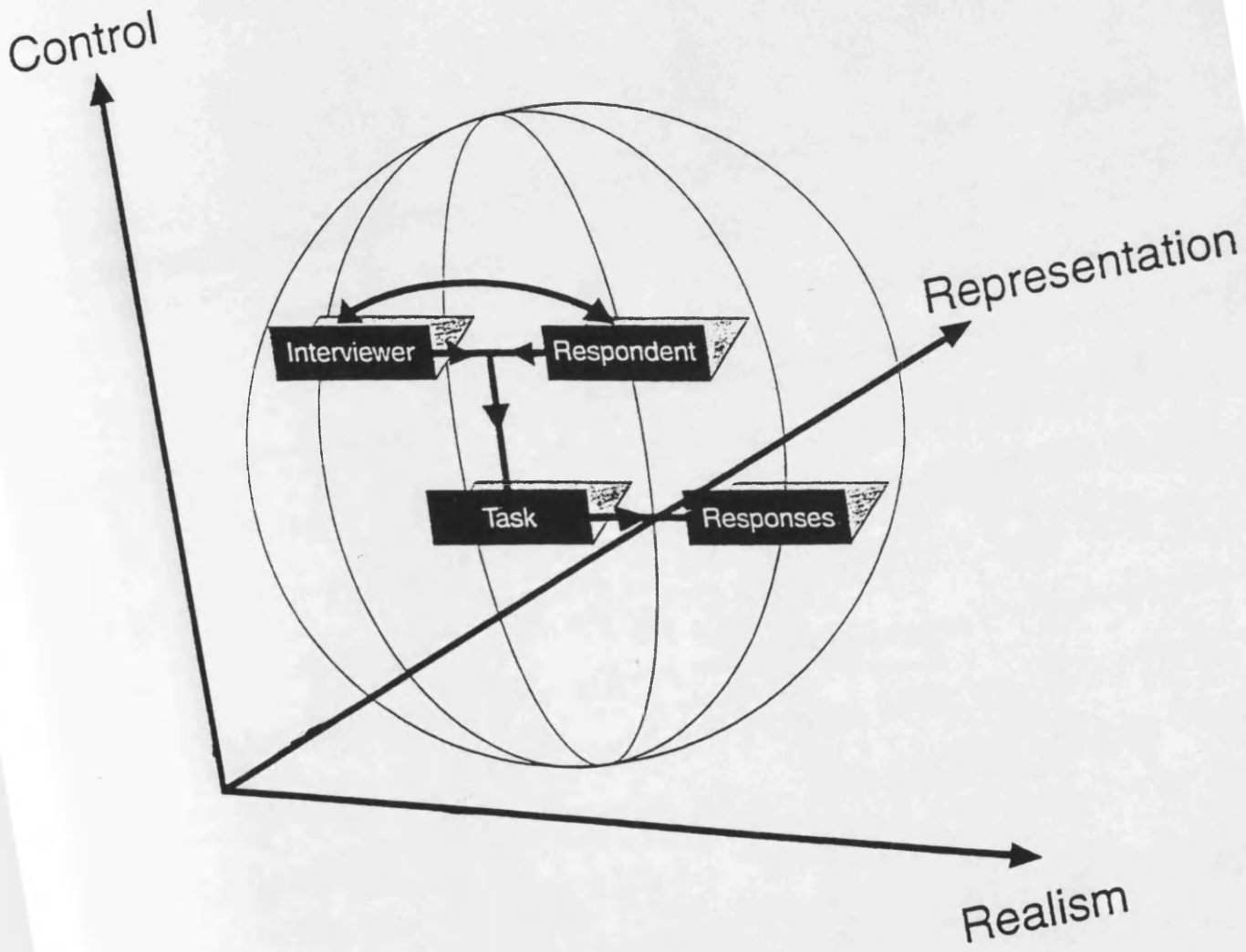


Figure 1.2 A framework for social research

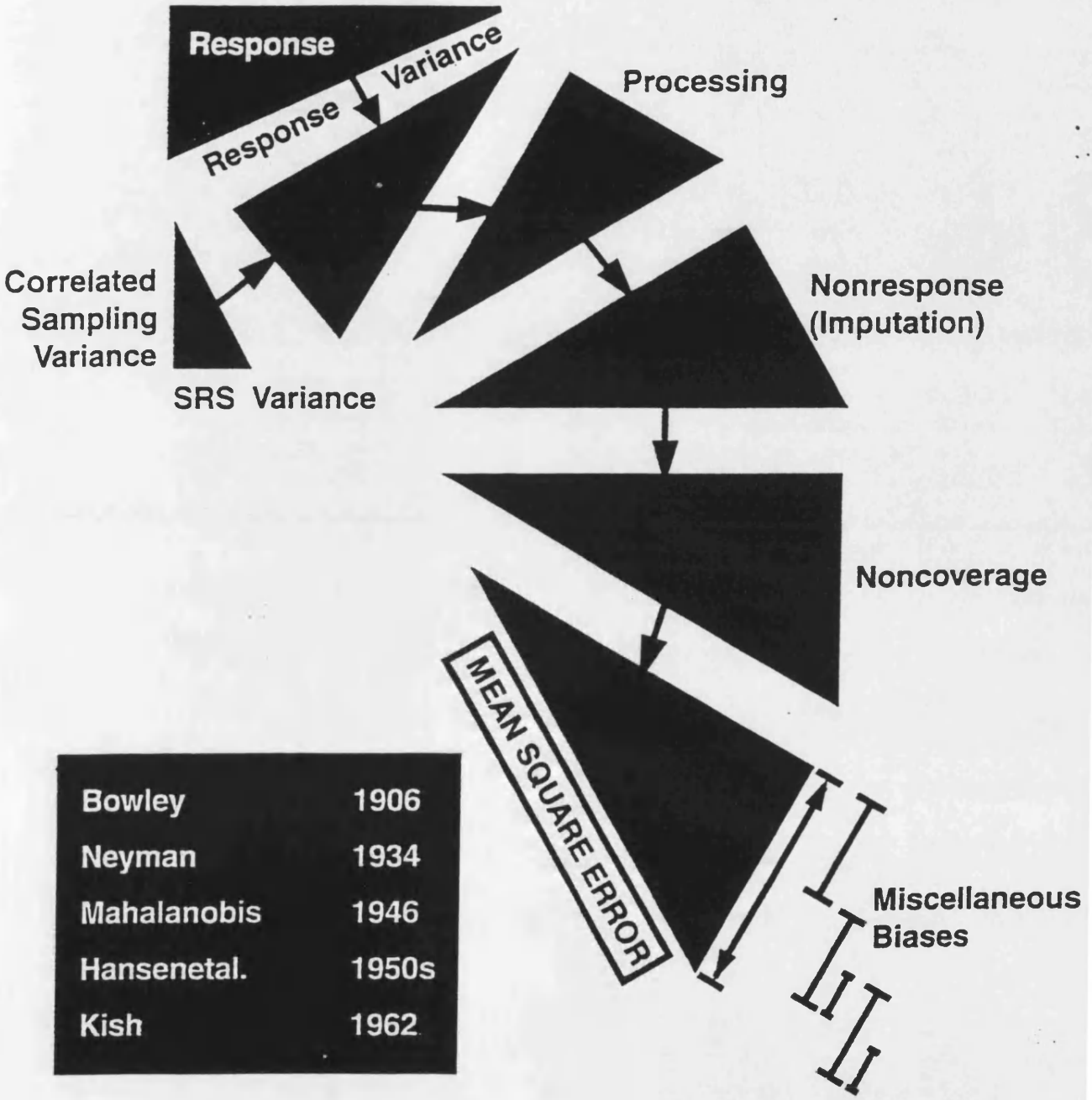


Figure 1.3 The *official statistics* model of survey error

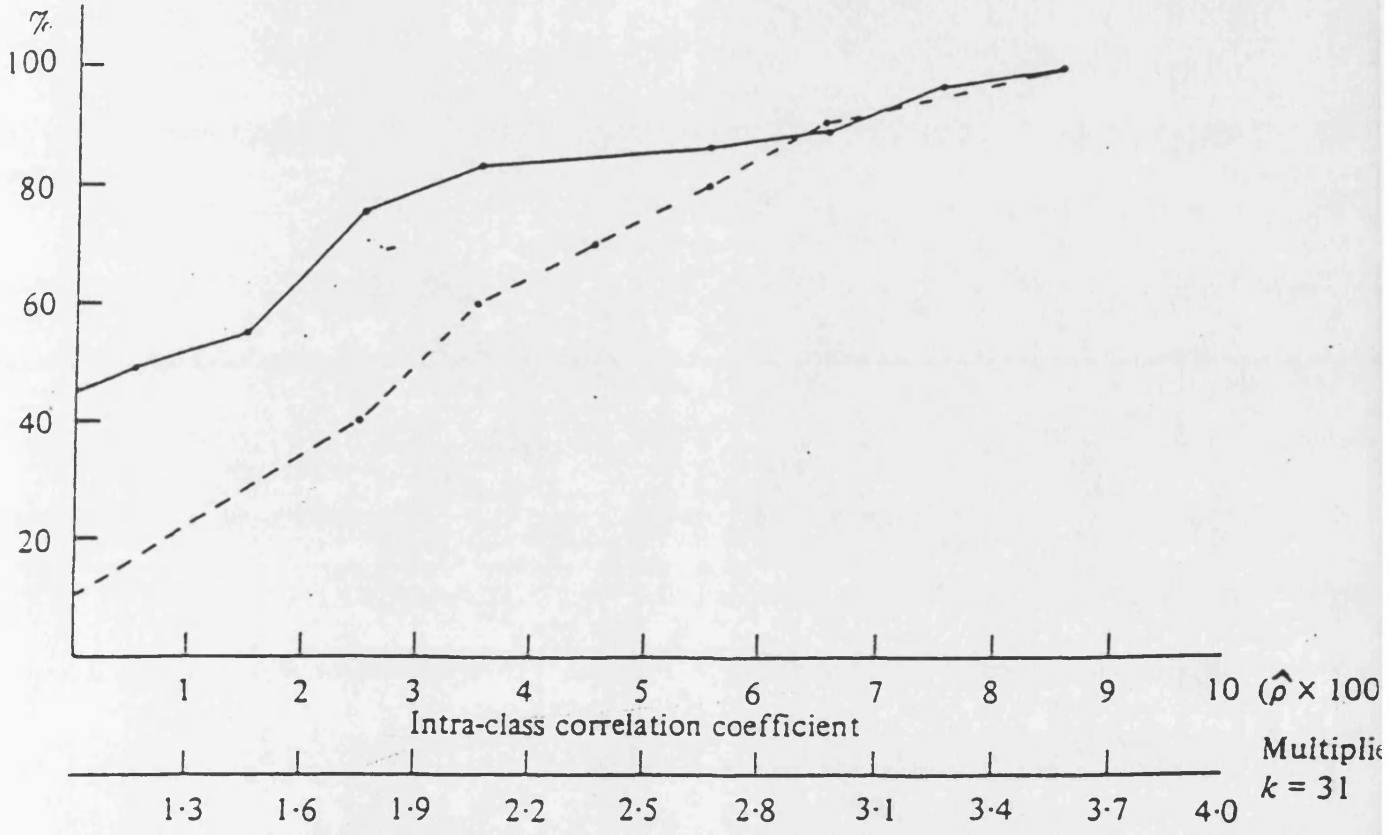
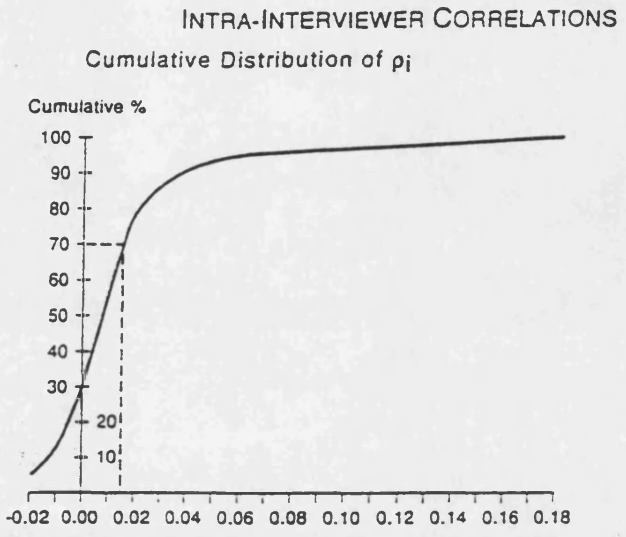


Figure 3.1 Cumulative ρ's for Noise Annoyance Survey



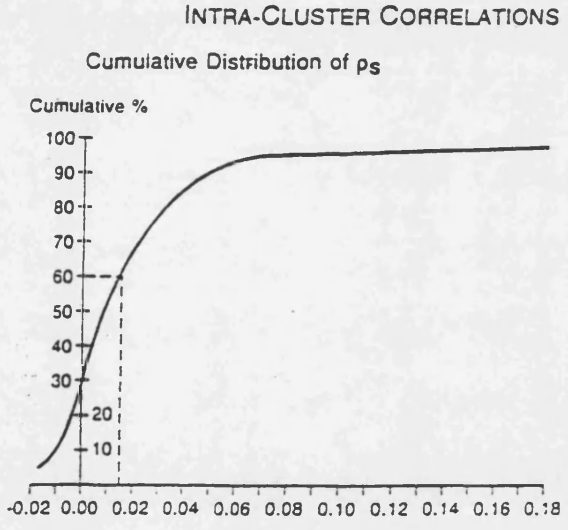


Figure 3.2 Cumulative ρ's for BHPS

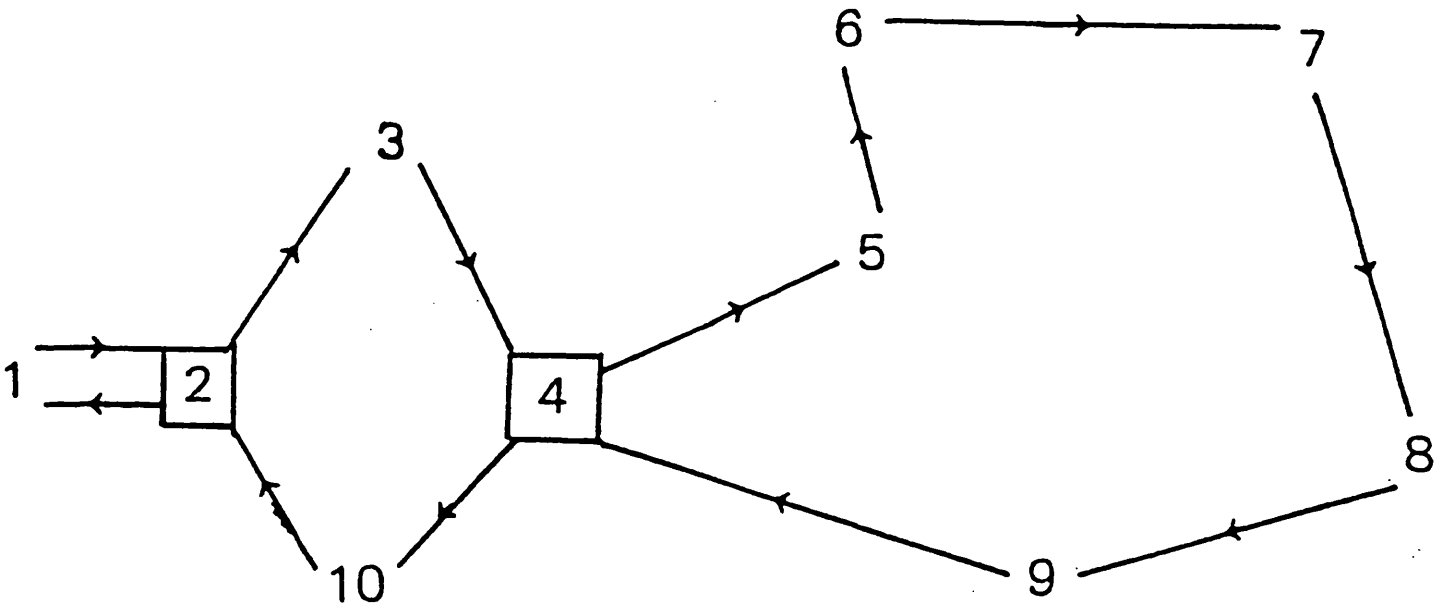


Figure 4.1 Pattern of fieldwork in Peru

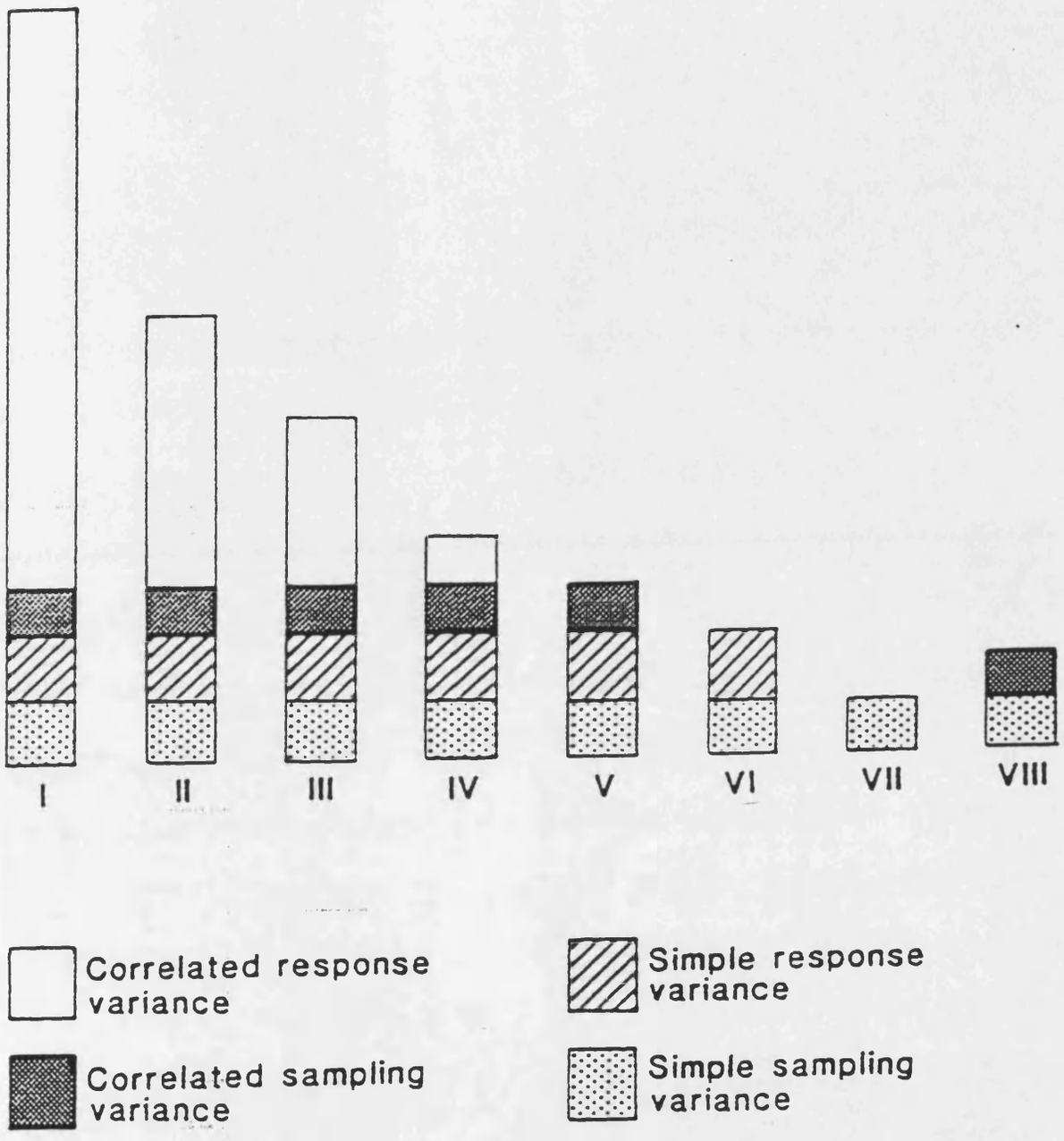


Figure 4.3L Lesotho total variance: *First birth interval*

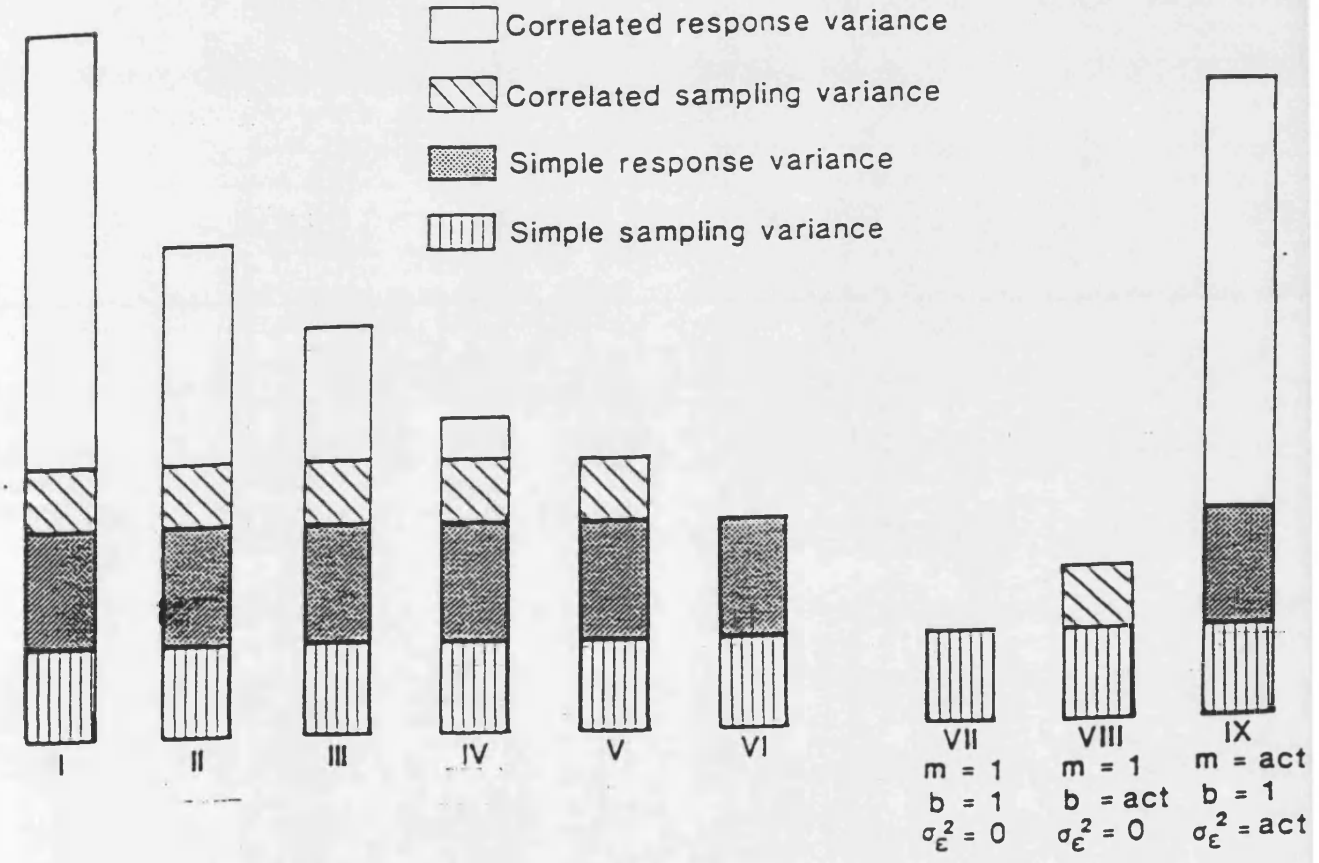


Figure 4.3P Peru total variance: ***First birth interval***

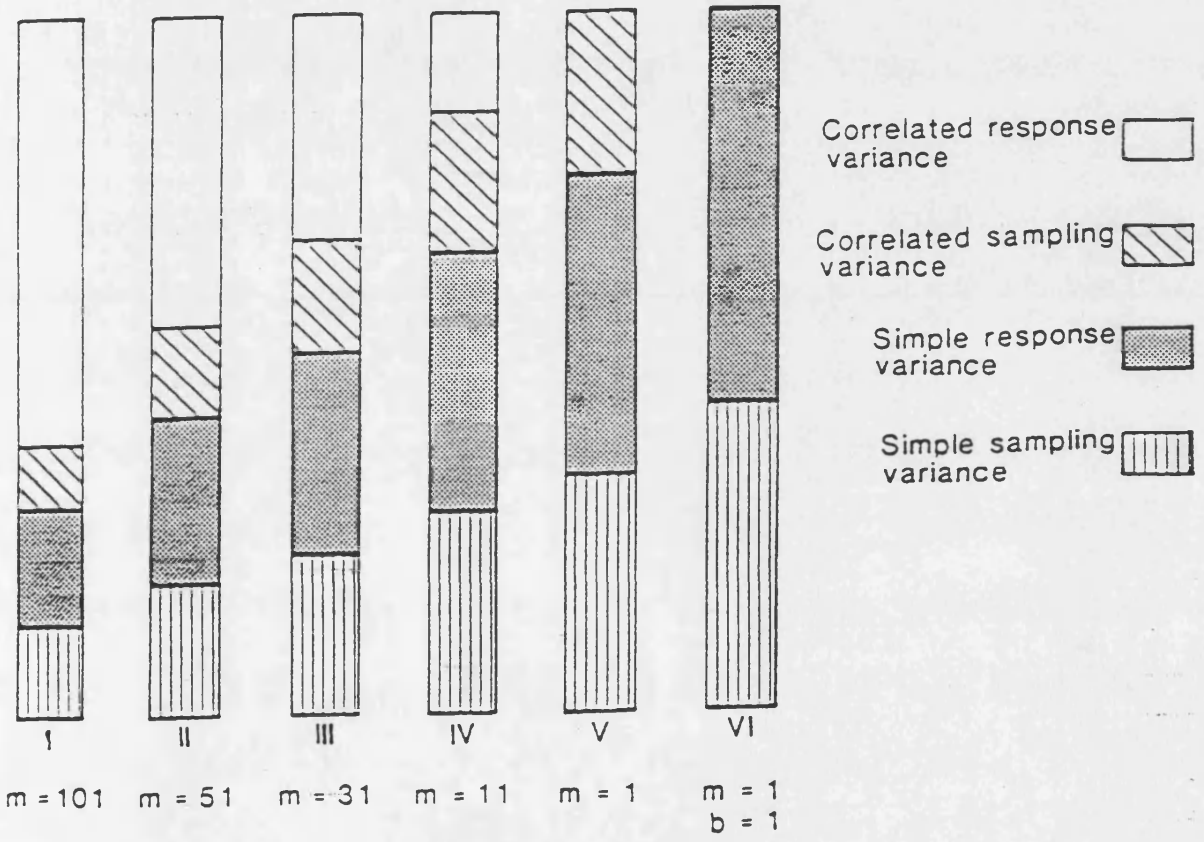


Figure 4.4P Peru total variance: *First birth interval* alternative presentation

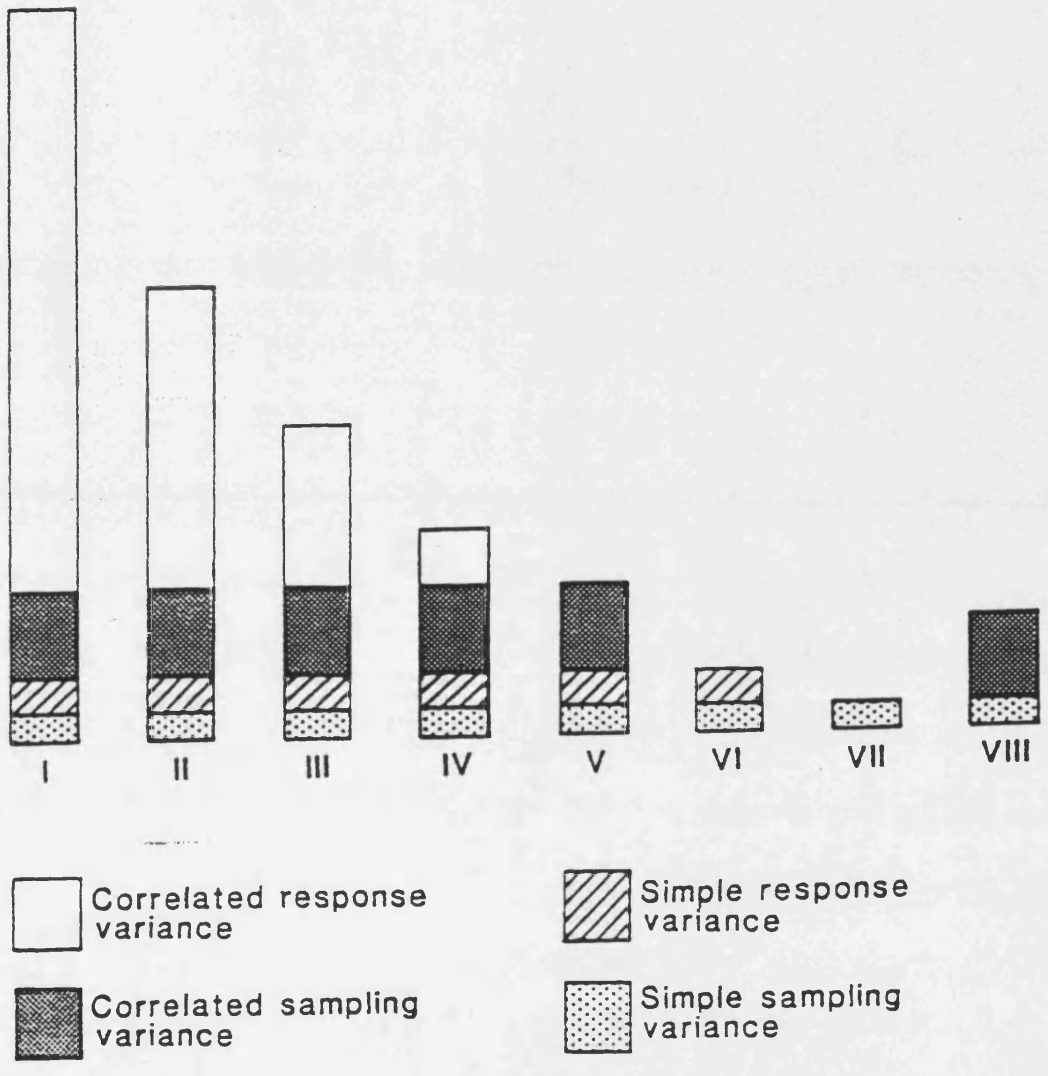


Figure 4.5L Lesotho total variance: *Ever-use of contraception*

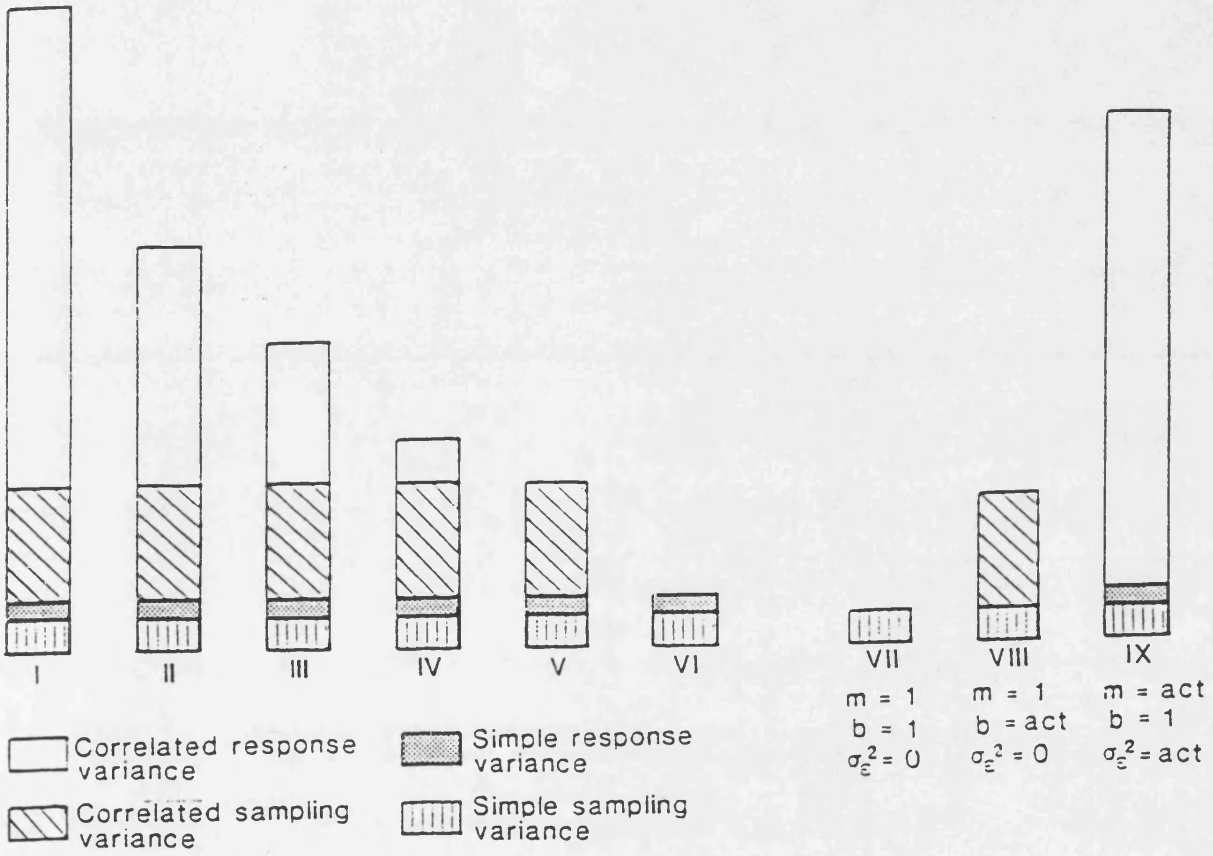


Figure 4.5P Peru total variance: ***Ever-use of contraception***

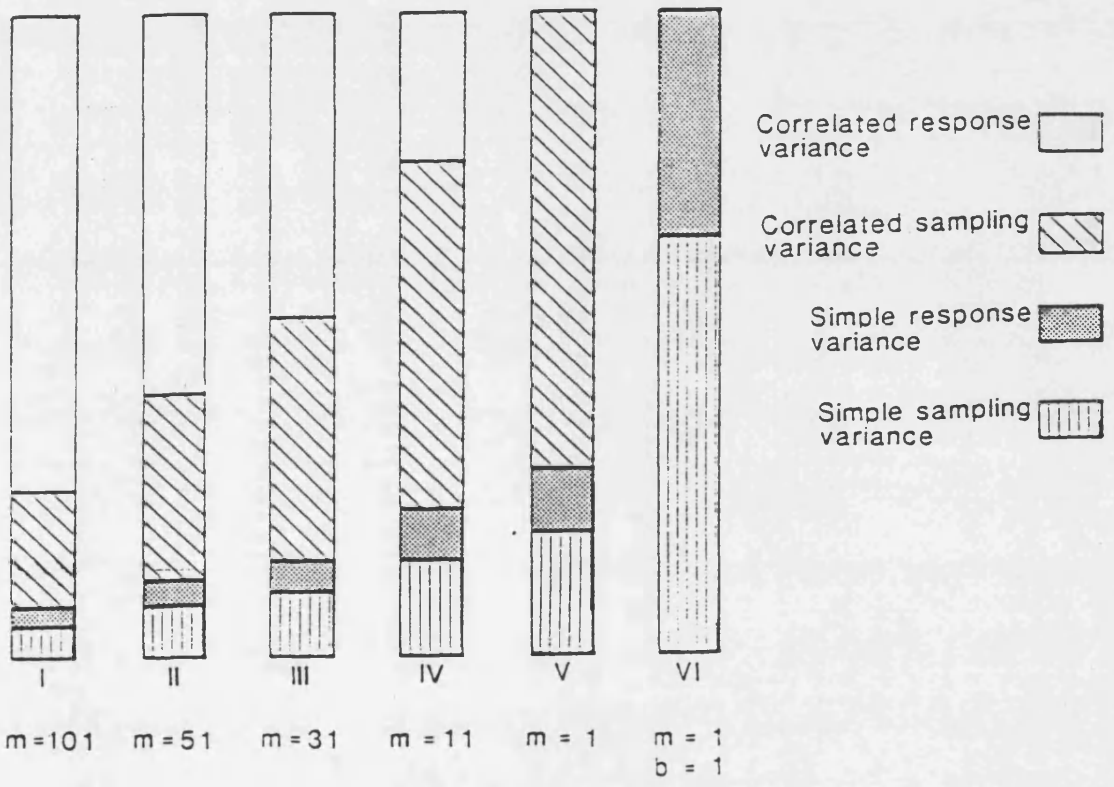


Figure 4.6P Peru total variance: *Ever-use of contraception* alternative presentation

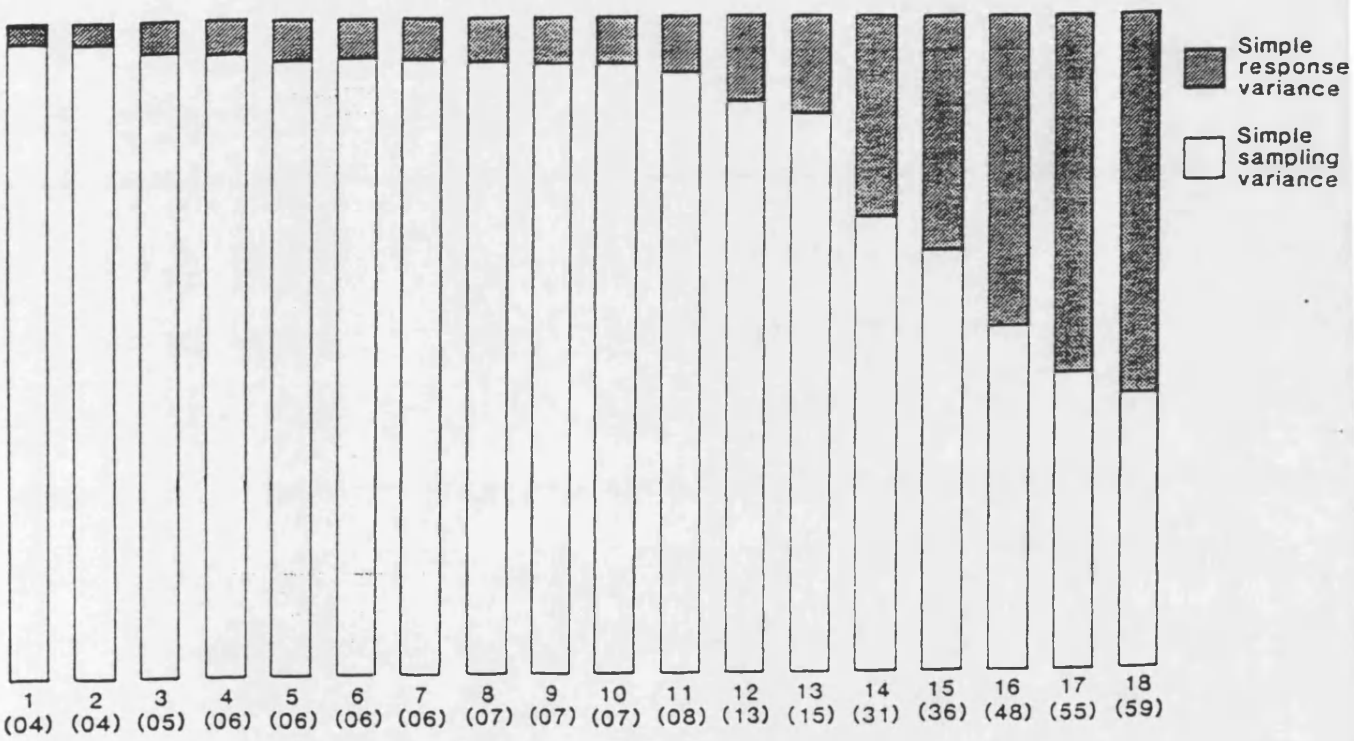


Figure 6.1L Components of STV for Lesotho

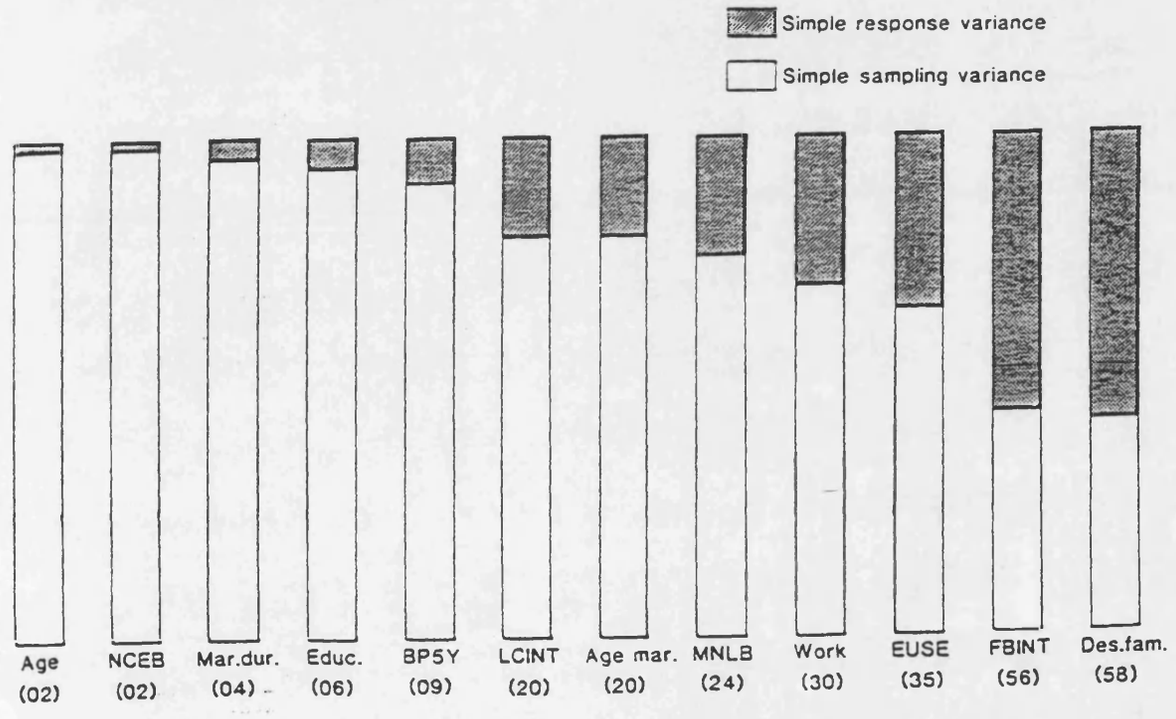
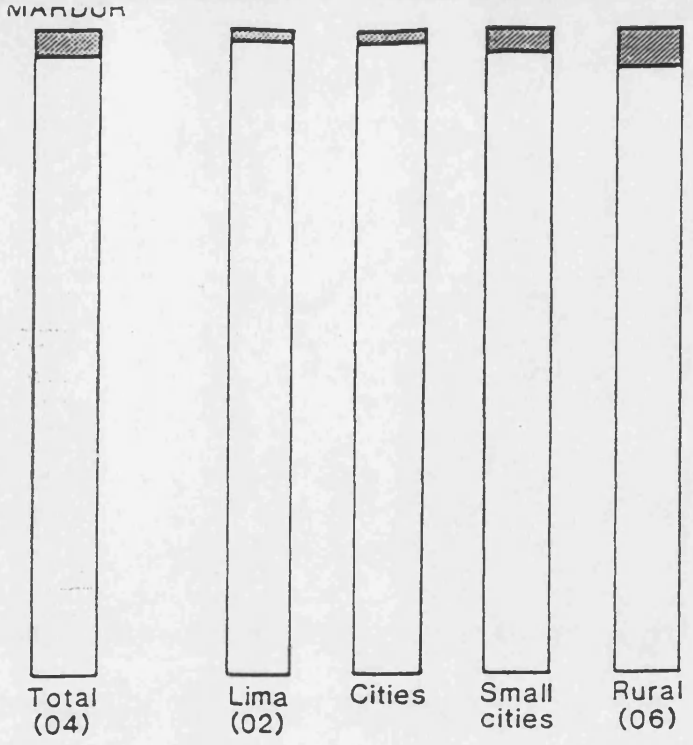


Figure 6.1P Components of STV for Peru



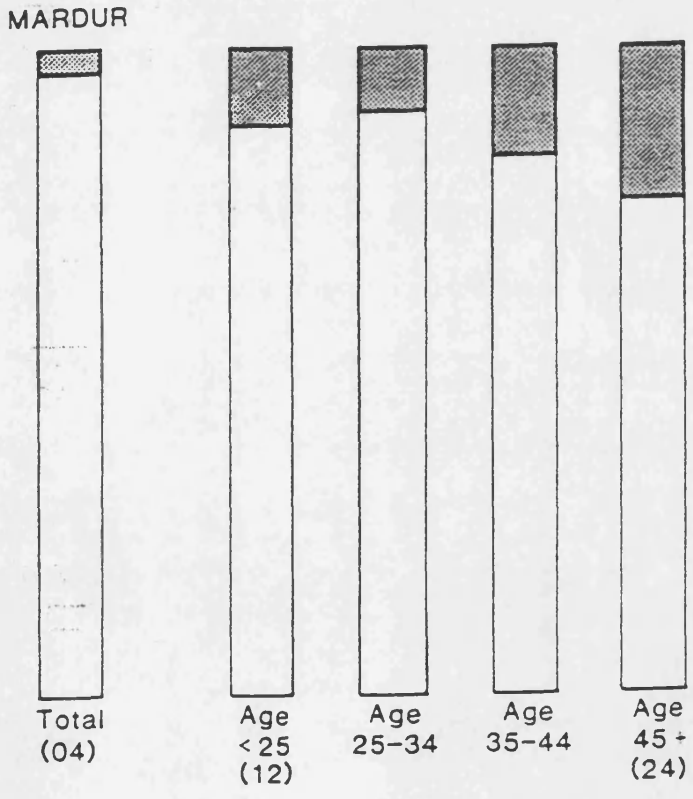
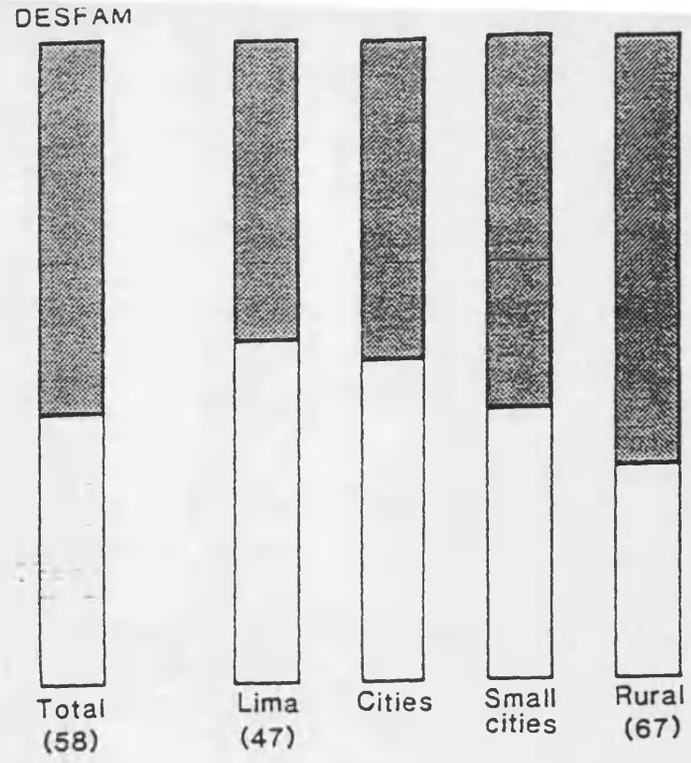


Figure 6.2 Partitioning of SRV for *Marital duration* for Peru



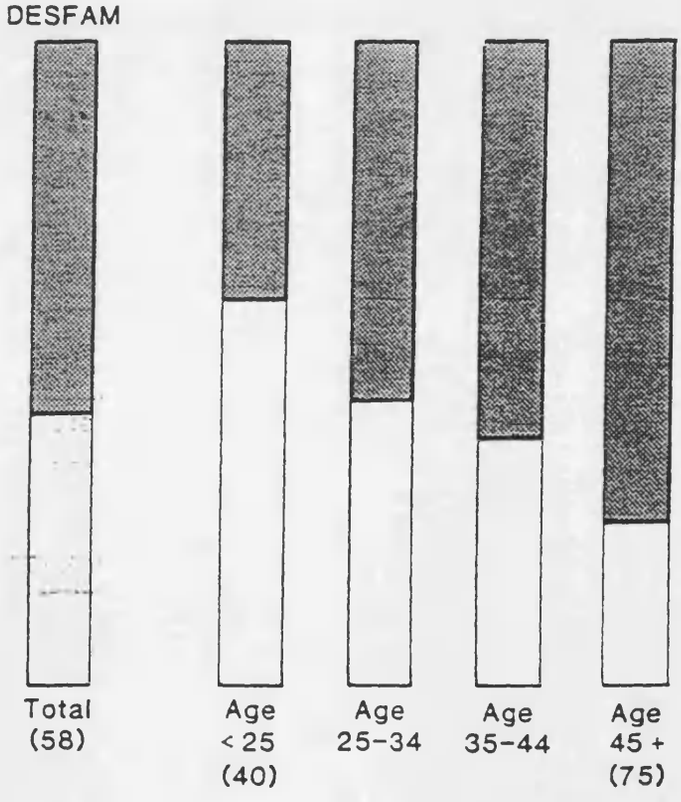


Figure 6.3 Partitioning of SRV for *Number of children desired* for Peru

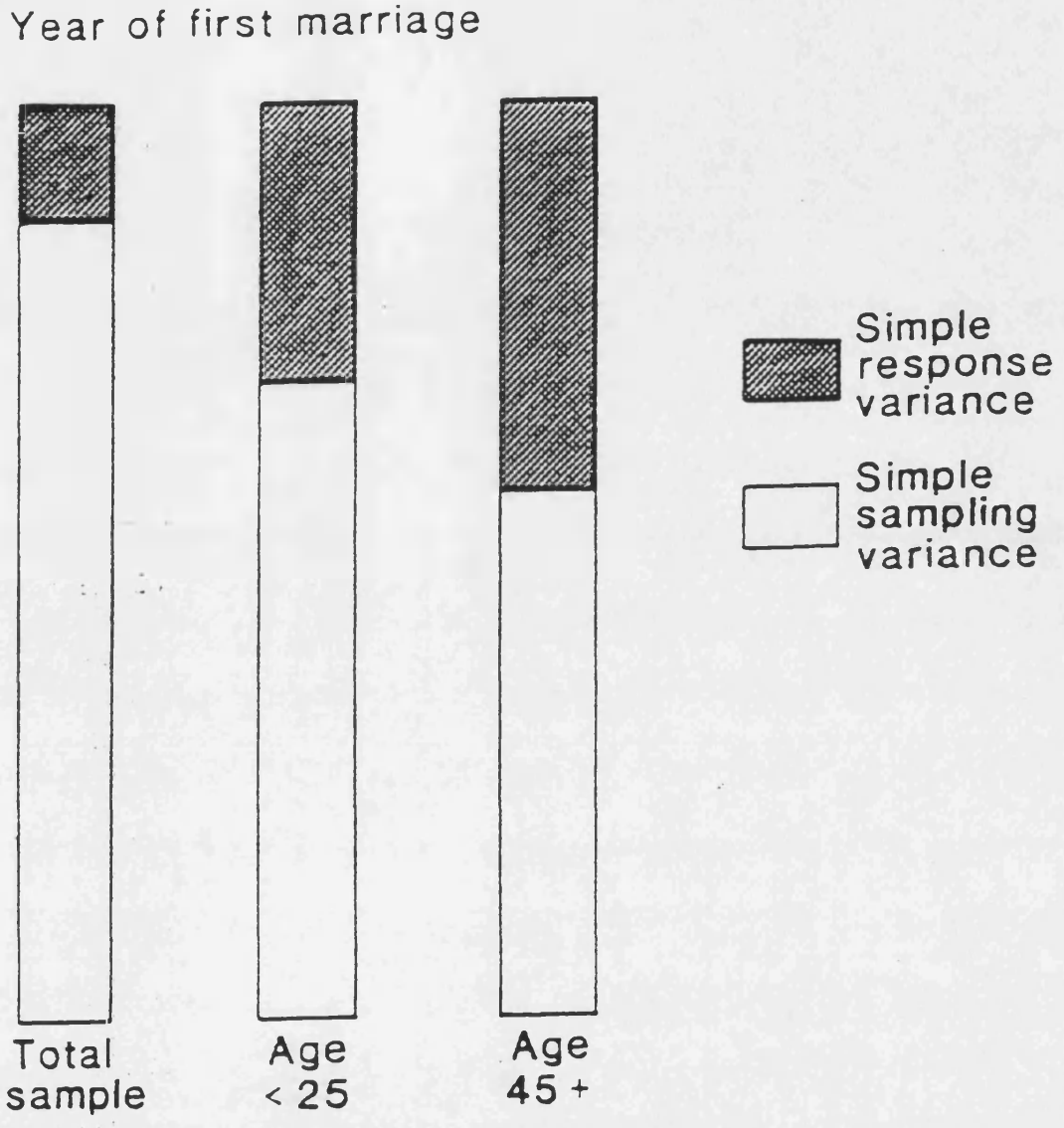


Figure 6.4 Interpretation of I for *Year of first marriage* in Lesotho

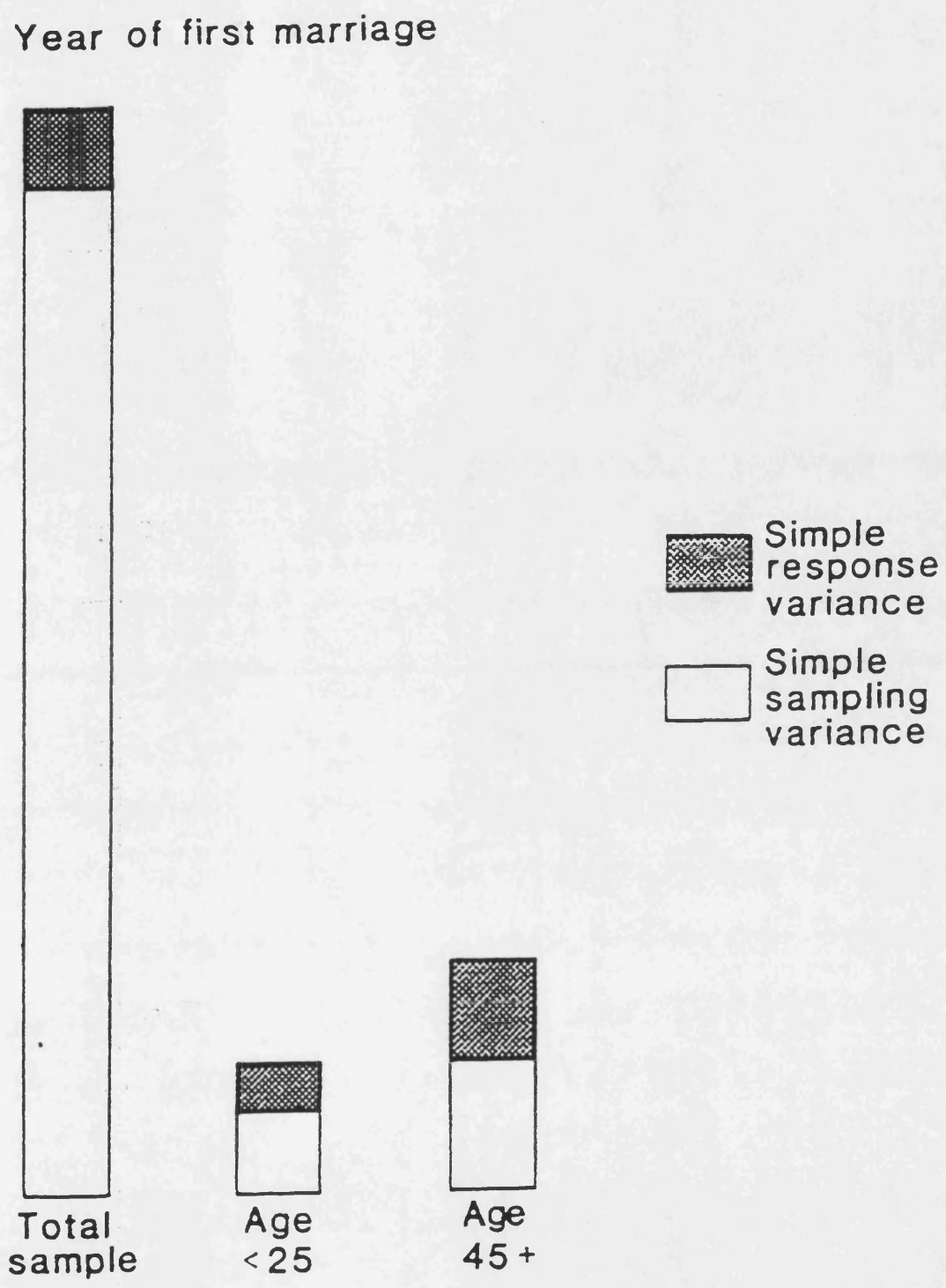


Figure 6.5 Contrast of I and SRV for age-related variables in Lesotho

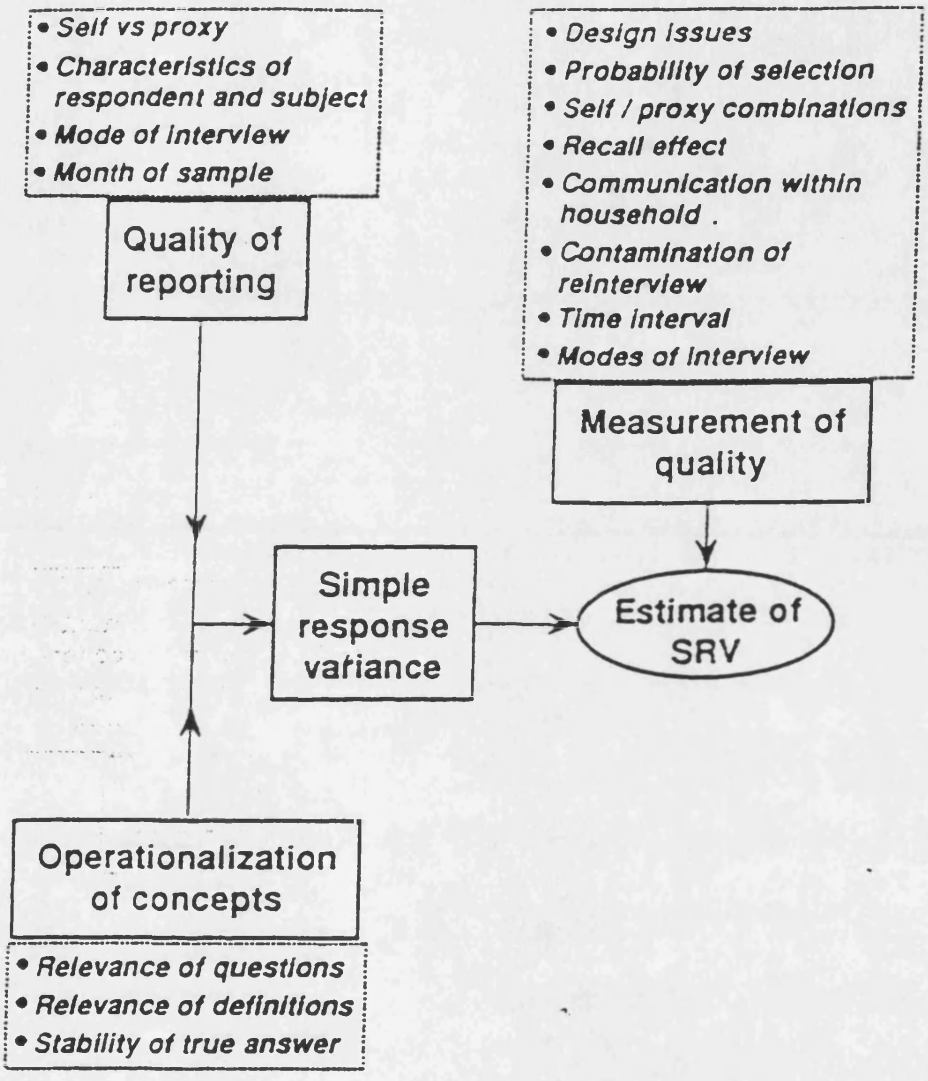


Figure 8.2 Factors affecting estimated SRV

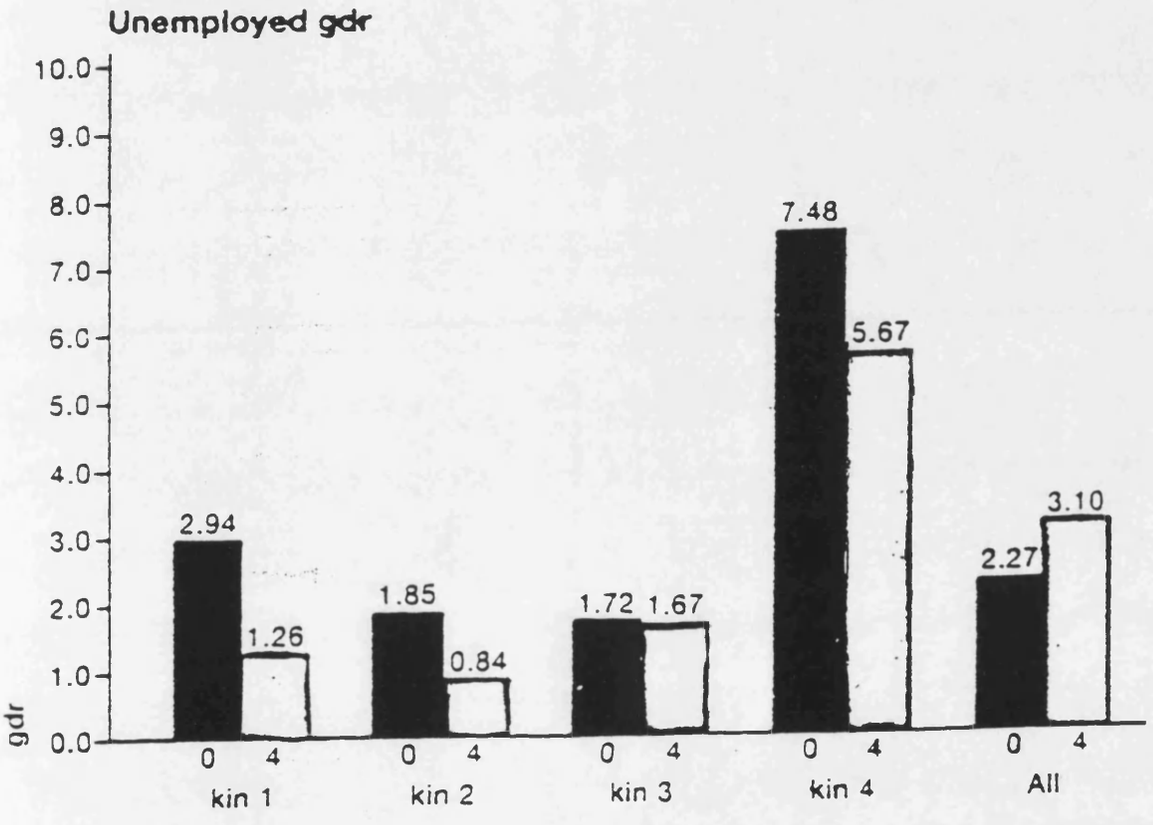


Figure 8.5A Simpson's paradox: comparison of self and proxy SRVs

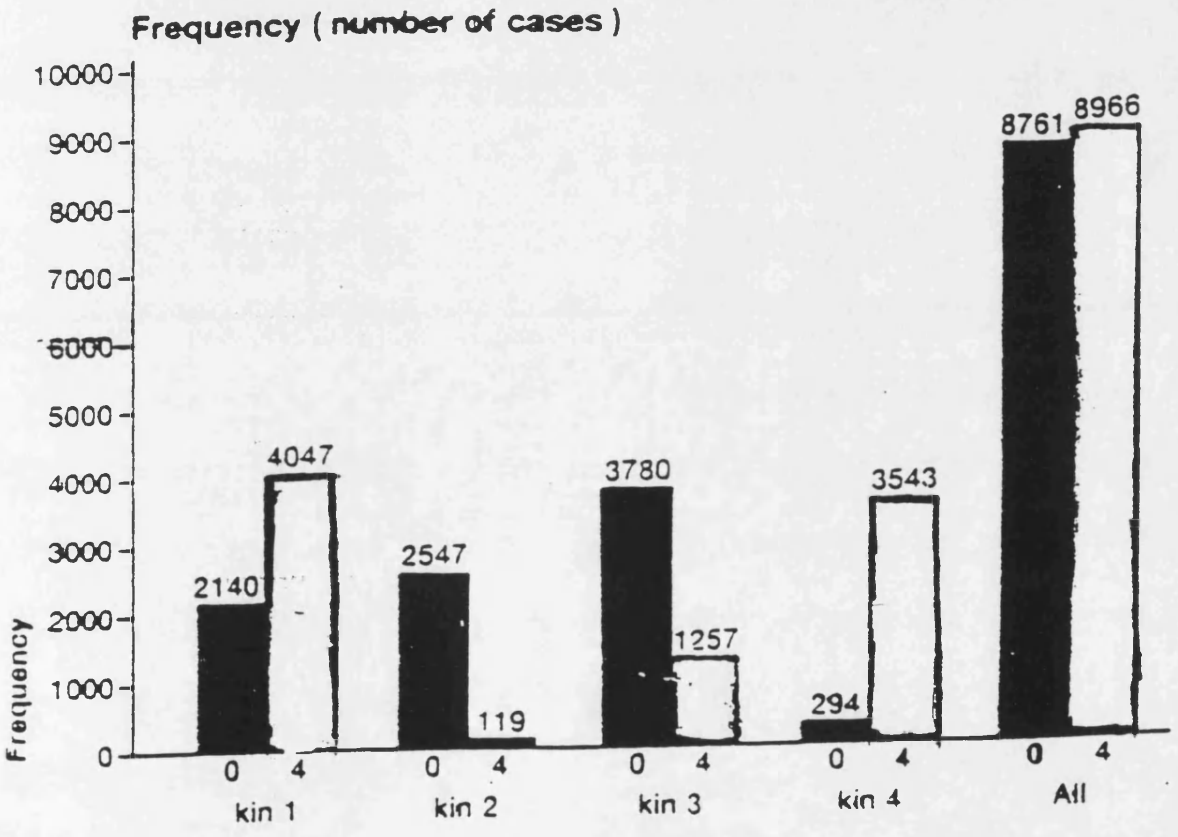


Figure 8.5B Simpson's paradox: distribution of respondents across categories

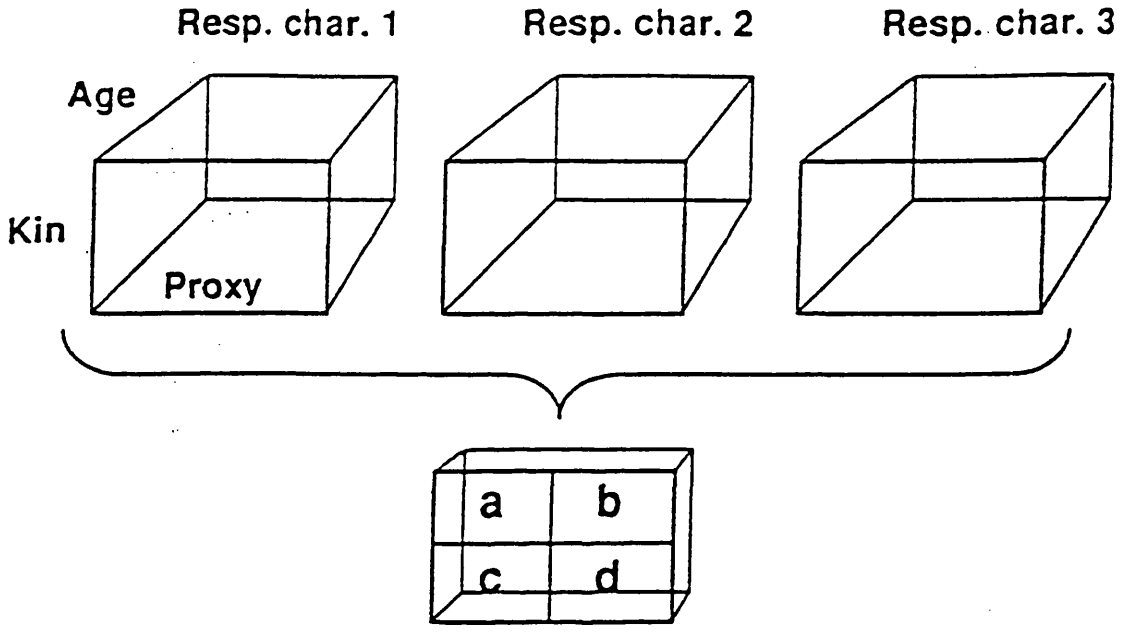


Figure 9.1 Interview-reinterview tables resulting from cross-classifying possible explanatory variables

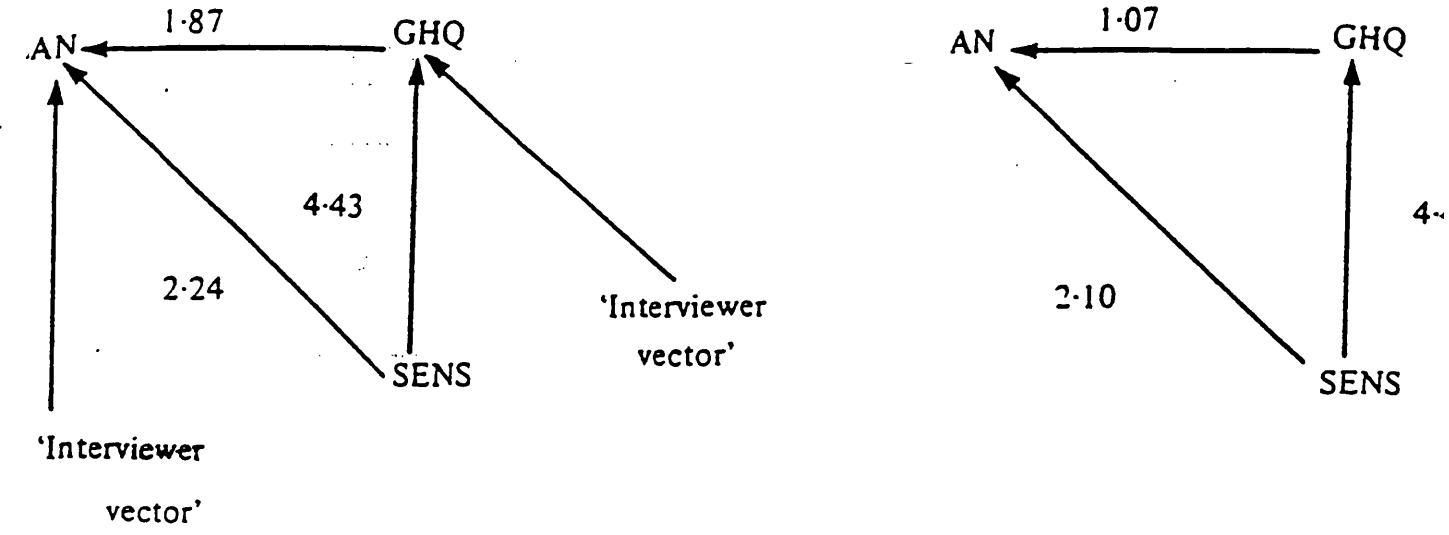


Figure 10.1 Path model for the effect of *Sensitivity* and *GHQ* score on *Annoyance*